UNIVERSIDADE NOVA DE LISBOA

Faculdade de Ciências e Tecnologia

Departamento de Engenharia Mecânica e Industrial

Implementação integrada de SPC e EPC na melhoria contínua do processo

Por

Octávio Ferreira Ramalho

Dissertação apresentada na faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa para obtenção do grau de Mestre em Engenharia Industrial

Orientador: Doutor José Fernando Gomes Requeijo

Lisboa

2009

Sumário

Durante os últimos anos tem havido nos meios académicos, por parte de um sector da engenharia industrial um esforço no sentido de integrar duas metodologias de controlo que, em muitas situações reais, estão implementadas de forma autónoma. A primeira, mais ligada à engenharia da qualidade, consiste na monitorização dos processos com recurso a cartas de controlo, ou controlo estatístico do processo (SPC - Statistical Process Control). A segunda abordagem está baseada no ajustamento dos processos recorrendo à informação sobre os níveis actuais das respostas ou desvios em relação aos valores de referência e constitui a engenharia de controlo do processo (EPC). Ambas as metodologias têm por base a identificação e modelação do processo.

O presente trabalho apresenta metodologias abrangentes para a identificação e modelação dos processos, baseadas em estimativas dos mínimos quadrados e de máxima verosimilhança, com o objectivo de se captar as dinâmicas internas dos processos e assim obter os modelos necessários à implementação integrada, ou não, do controlo estatístico (SPC) e engenharia de controlo (EPC - Engineering Process Control).

Pretendendo-se ser o mais abrangente possível, toda a abordagem foi feita para lidar com sistemas de múltiplas entradas e múltiplas saídas (MIMO Multi-Input/Multi-Output), considerando-se os outros modelos (SISO ou MISO) como casos particulares deste.

A metodologia de identificação implementada foi posteriormente utilizada com dados reais provenientes de um processo MISO, obtendo-se assim o modelo que serviu de base para simular o sistema de integrado SPC/EPC.

Com base nos resultados obtidos por simulação, pode-se concluir que o sistema integrado permite simultaneamente um melhor ajustamento do processo em relação aos valores de referência, reduzir a sua variabilidade, detectar alterações verificadas no processo e, caso seja possível, alterar o algoritmo de controlo de forma a responder às alterações verificadas.

Abstract

During the last years it has been having in the academic means, on the part of the industrial engineering an effort in the sense of integrating two control methodologies that, in many real situations, are implemented in independent way. The first of these, defended by the quality engineers, is statistical process monitoring by control charts, or statistical process control (SPC). The second approach is based on adjusting the process using information about its current level or deviation from the desired target. Both methodologies are based on model specification and model building of the process.

This work presents a comprehensive methodologies for both the model specification and model building of processes, based on least squares estimation and maximum likelihood estimation in order to capture the internal processes dynamics and thus to get the necessary models to be used on the integrated, whether or not, implementation of the statistical process control (SPC) and engineering process control (EPC).

Intending to be so comprehensive as possible, the whole approach was made to lead systems of multiple input, multiple output (MIMO), considering the other models (SISO or MISO) as special cases of this.

The methods of model specification and model building implemented were later used with real data from a MISO process, thus obtaining the model that formed the basis for simulating the system of integrated SPC / EPC.

Based on the results of the simulation, we conclude that the integrated system allows both a better adjustment process in relation to reference values (target), reduce their variability, detecting changes in process and, if possible, to change the algorithm control to respond to process shifts.

Indice

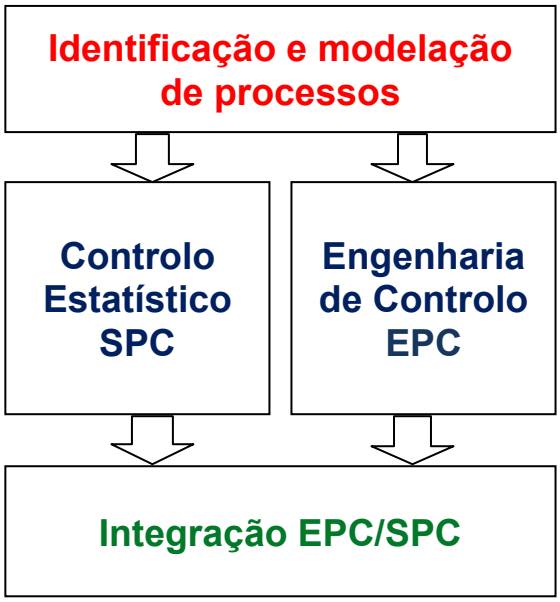
1 Enquadramento

O desempenho de um produto, um sistema ou um serviço normalmente é julgado em termos de confiança (que pode ser definido como um agregado de qualidade, fiabilidade, manutibilidade etc.) e segurança, não negligenciando o custo de alcançar estes atributos.

O controlo eficiente dos processos é um elemento chave na manutenção e melhoria da qualidade e produtividade. Tradicionalmente, duas diferentes vertentes técnicas têm contribuído significativamente para o desenvolvimento do controlo dos processos na indústria

As técnicas de ajustamento do processo baseadas em princípios de controlo por realimentação ou *feedback* tornaram-se um importante recurso do kit de ferramentas utilizado pelos Engenheiros da Qualidade (ver (Box & Luceño, 1997), (Montgomery, 2001), (Del Castillo E. , 2002) , (Box, Jenkins, & Reinsel, 2008)). Uma variedade de técnicas para o ajustamento dos processos tem sido propostas e estudadas devido ao recente interesse na integração do controlo estatístico do processo (*SPC – Statistical Process Control*) e controlo da engenharia do processo (EPC – Engineering Process Control).

Este tema envolve a abordagem de quatro grandes áreas de estudo: identificação e modelação de sistemas ou processos, controlo estatístico, engenharia de controlo e a integração do controlo estatístico com a engenharia de controlo.



A parte fulcral da agregação destas quatro áreas, e à qual foi dedicado a maioria do tempo durante este trabalho, é a parte de identificação e modelação sistemas. A obtenção de um bom modelo é meio caminho andado para a obtenção de bons resultados na aplicação desta metodologia.

A engenharia de controlo tem por base o modelo do processo. Apesar de a maioria das metodologias de controlo, P, PI, PID, utilizarem o modelo off-line como uma ferramenta de análise e projecto, no controlo preditivo, o modelo é a parte integrante do algoritmo de controlo e os níveis resultantes são aplicados directamente no processo.

No controlo estatístico, na utilização das cartas de controlo assume-se que, se o processo está sob controlo, o processo tem média constante e os dados são completamente não correlacionados. Em engenharia de processo defende-se que quase todos os processos produtivos exibem autocorrelação (Del Castillo E. , 2002), pelo que a obtenção de um bom modelo para o processo pressupõe que os resíduos provenientes do modelo sejam constituídos unicamente por ruído branco, alcançando-se assim as condições ideais para a implementação do controlo estatístico.

1.1 Aspectos de controlo do processo

O termo controlo do processo é utilizado de diferentes modos. As cartas de Shewhart e outras cartas controlo da qualidade são frequentemente empregues na indústria, maioritariamente de produção discreta, recorrendo a medidas parciais da produção no que é chamado controlo estatístico do processo (SPC). Por contraste, particularmente na indústria de processo e química, utilizam-se várias formas de ajustamento por feedback ou feedfoward no que é chamado controlo de engenharia do processo (EPC). Como os ajustamentos feitos são normalmente calculados por computador e aplicados automaticamente, este tipo de controlo também é conhecido como controlo automático do processo (APC).

1.1.1 Monitorização do processo e ajustamento do processo

O controlo do processo não é menos que tentar cancelar/remover os efeitos de uma lei fundamental da física, a segunda lei da termodinâmica, que diz que, se não houver intervenção externa, a entropia ou desorganização de qualquer sistema nunca diminui e normalmente aumenta (Box, Jenkins, & Reinsel, 2008). O SPC e o EPC são duas abordagens complementares para combater esta lei. O SPC tenta remover os distúrbios através da monitorização do processo enquanto que o EPC tenta compensá-los através do ajustamento do processo.

1.1.1.1 Monitorização do processo

A estratégia do SPC para estabilização dos processos consiste em estandardizar procedimentos e matérias-primas e utilizar dispositivos geradores de hipóteses (tais com gráficos, listas de comprovação, cartas de Pareto, diagramas causa-efeito, etc.) para seguir e eliminar causas de problemas. Uma vez que a procura de causas atribuídas é caro e tedioso, normalmente opta-se por esperar até que ocorram desvios estatisticamente significantes para se instituir essa procura. Para implementar este processo recorre-se a cartas de monitorização do processo tais como cartas de Shewhart, cartas Cusum, cartas EWMA etc.

1.1.1.2 Ajustamento do processo

Apesar do grande empenho dedicado em remover as causas de variação, tais como métodos de testes não satisfatórios, diferenças na matéria-prima, diferenças de operador

etc., em muitos dos processos não é possível estabilizá-los num estado satisfatório apenas recorrendo a estas técnicas. Apesar dos esforços, o processo continua a exibir a tendência de se afastar dos níveis pretendidos. Estas situações podem verificar-se devido a fenómenos conhecidos mas incontroláveis como a variação da temperatura ambiente, humidade, etc. ou correntemente não conhecidas. Nestas circunstâncias, será necessário recorrer a sistemas de ajustamento ou regulação onde a manipulação de determinadas variáveis de entrada do processo tem o efeito de compensar as perturbações que afectam as características da qualidade do processo.

.

2 Objectivos

Sem haver um objectivo previamente definido na abordagem deste tema, tentou-se, dentro das restrições impostas ao desenvolvimento deste trabalho, ir o mais longe possível no estabelecimento, organização e implementação de ferramentas que permitissem construir uma metodologia mais ou menos abrangente para o desenvolvimento de técnicas para a implementação de sistemas integrados EPC/SPC em ambiente industrial. Com o andar do tempo foi-se tendo a sensibilidade da dimensão da matéria a abordar e a ficar com a sensação de que seria difícil sair dos fundamentos sem queimar etapas, optando-se por "alargar" as bases em vez de se avançar para modelos de controlo mais elaborado e específicos.

Sendo um dos objectivos do trabalho a estruturação da metodologia, tentou-se abdicar das ferramentas que por vezes se utilizam sem se saber ao certo o que se está a fazer, pelo que, na modelação de sistemas dinâmicos, e noutros desenvolvimentos usados ao longo deste trabalho, evitou-se ao máximo utilizar software existente, como por exemplo MATLAB'S SYSTEMS ID TOOLBOX ou SAS PROC STATE SPACE, uma vez que ao recorrer-se a este tipo de ferramentas, a metodologia proposta passa a depender de ferramentas externas não generalizadas, para além de se perder a teoria que está por subjacente a esta metodologia, perdendo-se assim também a sensibilidade de análise de dos resultados.

Devido à grande extensão da matéria envolvida neste tema, e às restrições em termos de tempo, houve muitos conceitos e metodologias que foram deliberadamente postos de parte ou abordadas de forma muito ligeira, como é o caso da metodologia de espaço de estados, controladores PID, controladores PID com modelo interno, controlo preditivo, forma canónica do modelo VARMA ou mesmo os modelos não estacionários e de rank reduzido, sendo que alguns assuntos foram mesmo utilizados na prática, nas não estão devidamente documentados. Devido ao seu inquestionável interesse, esses temas deverão ser desenvolvidos posteriormente no seguimento deste trabalho.

O suporte base de programação e implementação das fórmulas e algoritmos apresentados ao longo da parte teórica foi basicamente o MATLAB, versão 7.0.0.19920 Da Math Soft Inc.

Em certos casos recorreu-se ao Microsoft Visual C++, versão 6, principalmente quando se pretendeu criar ferramentas para utilização futura como foi o caso dos programas para a determinação de custos mínimo como os apresentados nos anexos I e II. Estas ferramentas foram desenvolvidas numa altura em que os objectivos e processos de estudo ainda não estavam definidos com a intuição de que deveriam ser necessárias numa altura mais avançada.

3 Introdução

3.1.1 Controlo do processo em indústrias de processo e produto

Nas industrias de base de produto o objectivo é manter as características da qualidade tão próximas quanto possíveis dos valores de referência ou *target.* A conformidade exacta com os valores de referência só poderá ser atingível em termos teóricos uma vez que existem uma grande variedade de factores que afectam os processos de produção causando desvios em relação aos valores desejados. O objectivo pode ser atingido por processos estatísticos de controlo (*SPC- Statiatical Process Control*), cuja ferramenta principal passa pelos registos e interpretação das cartas de controlo. As características do processo tais como a média, e respectivos desvios, de um processo contínuo, não conformidades ou percentagem de não conformidade são monitorizadas numa carta de amostragens do processo ao longo do tempo. Enquanto os registos caírem dentro dos limites de controlo e não violarem certas regras, entretanto impostas para melhorar o desempenho das cartas, não haverá qualquer interferência no processo. Sempre que os registos caiam fora dos limites de controlo ou se verifique a violação de uma das regras impostas, então procuram-se as causas que provocaram a alteração do processo. A esse tipo de causas atribuiu-se o nome de causas atribuídas, ou causas designáveis. O controlo estatístico tem assim uma visão binária das condições do processo, está a correr em condições satisfatórias ou não. O objectivo básico é diferenciar entre causas aleatórias inevitáveis, chamadas causas comuns, e causas atribuídas no processo e a fronteira entre ambas está basicamente nos limites de controlo. Se apenas se verificarem causas comuns, o processo continua. Se for detectada uma causa atribuída, o processo pára para se detectar e eliminar a causa. Ferramentas do controlo estatístico (SPC) tais como as cartas de controlo de Shewart, cartas de média móvel exponencialmente amortecidas (*EWMA – Exponential Weighted Moving Average*) e cartas de somas acumuladas são normalmente utilizadas com estes objectivos.

A engenharia de controlo do processo (*EPC - Engineering Process Control*) tem sido utilizado em processos de produção continua. O EPC é constituído por uma colecção de técnicas para manipular as variáveis ajustáveis do processo com o objectivo de manter e/ou conduzir a saída do processo o mais próximo possível dos valores desejáveis. O objectivo do EPC é providenciar uma resposta instantânea, contrabalançando as alterações no balanço do processo e aplicar acções correctivas para trazer a saída para os valores alvo. A abordagem consiste em prever os desvios da saída que possam ocorrer caso não exista acções de controlo e actuar de modo a cancelar os desvios previstos. Para atingir os objectivos EPC recorre-se a apropriados algoritmos de controlo por realimentação (*feedback*) ou controlo em avanço (*feedforward*) que indicam quando e por quanto é que o processo deve de ser ajustado para atingir os objectivos.

O primeiro passo no ajustamento por feedback é construir um modelo preditivo para o processo, determinando como é que as saídas e as entradas estão relacionadas. Esta tarefa é das mais importantes e é a base para uma boa estratégia de ajustamento. Para a obtenção do modelo, inicialmente recorria-se a metodologias como desenho de experiencias e resposta em superfície, essa tarefa era normalmente executada off-line.

Estas ferramentas revelaram-se bastante úteis na presença de processos responsivos, nos quais o comportamento dinâmico das variáveis de saída é apenas devido à dinâmica dos distúrbios e o controlo exercido tem seu efeito completo imediatamente. Na produção discreta, o factor de controlo será o ponto de operação da máquina (*set point*). Ás alterações na saída no estado estacionário que se obtém devido a uma alteração unitária na entrada dá-se o nome de ganho do sistema. O valor do ganho é determinado off-line com recurso a desenho de experiências e a técnicas de regressão. A literatura disponível em processos de ajustamento pode ser amplamente classificada de acordo com os problemas a que se destinam:

1. Feedback adjustment for machine tool problems
2. Setup adjustment problems
3. Run-to-run process control in application to semiconductor industry

3.1.1.1 A necessidade do complemento EPC-SPC

Apesar do SPC e EPC terem sido desenvolvidos em campos diferentes para os respectivos objectivos, estas metodologias podem se complementar na obtenção do objectivo comum de redução da variabilidade. Os seguintes pontos destacam a necessidade de ajustamento dos processos na indústria:

1. Os ambientes produtivos são não estacionários e os processos estão sujeitos a alterações ocasionais. Apesar das causas das alterações serem conhecidas, estas podem ser impossíveis ou economicamente inviáveis. Alguns exemplos são a variabilidade da matéria-prima, alterações nos processos devido à manutenção, variações na temperatura ambiente e humidade, etc. Tais fontes de variabilidade são inevitáveis e geralmente não podem ser eliminadas apenas pela monitorização do processo. Deve-se então recorrer ao ajustamento do processo para compensar e minimizar a variabilidade em tais circunstâncias.
2. O processo pode sofrer deriva lenta. A deriva pode dever-se a causas conhecidas tais como a acumulação de resíduos dentro de um reactor, o envelhecimento de componentes, etc. que podem não ser identificadas com precisão. A utilização apenas do SPC não será bem apropriada para controlo de processos com deriva lenta. Sem interferência no controlo do processo, o processo pode atingir determinados desvios antes que qualquer acção de controlo seja tomada como resposta a um alarme. Se o custo devido ao produto fora de especificação é elevado ou os custos de ajustamento baixos, não há necessidade de esperar um longo período para se observar um ponto fora dos limites e disparar uma acção de controlo.
3. Existem processos em que o estado de controlo estatístico pode tratar-se de um caso ideal e portanto difícil de atingir ou dizer que o processo está sob controlo estatístico. Nesses casos seria benéfico ter um controlo moderado com ajustamento do processo.
4. O ajustamento do processo por si só, não é apropriado para eliminar causas especiais que podem afectar o processo. Quando acontecem causas especiais, tais como alterações súbitas das condições ambientais ou erros de leitura, entre outras, o ajustamento do processo por si só não lidará com tais situações. Estas causas provocarão situações de desvios em relação ao

target e o aumento da variabilidade da saída. Estas causas devem ser detectadas com recurso à monitorização do processo.

3.1.1.2 Argumentos contra o ajustamento do processo

Num passado recente, os Estatísticos e os Engenheiros de processo aderiram à noção de "não interferir no processo se ele está sob controlo estatístico". Evitou-se a ideia do ajustamento do processo. Esta noção, então defendida por Deming através da experiência popularmente conhecida do funil de Deming. Deming estudou os efeitos de não ajustamento e ajustamento na variância do processo. Ele conclui que a estratégia de não ajustamento produzia melhores resultados e o processo mantinha-se nos valores alvos. Procedendo-se a uma análise mais profunda da experiência compreende-se os pressupostos da experiência que conduziram a essa conclusão:

1. O processo fonte dos desvios está sob controlo estatístico.
2. O processo está inicialmente no alvo.
3. É possível apontar o funil ao alvo.

A mesma experiência foi mais tarde analisada e tiraram-se algumas conclusões úteis. Quando o processo está fixo no alvo e sob controlo estatístico não precisa de ajustamento. Contudo, se um processo não controlado exibe autocorrelação, o recurso a regras de controlo poderá melhorar significativamente o seu desempenho. Se o processo for não estacionário, a própria média move-se, pelo que se o processo não for controlado, a média afastar-se-á dos valores alvo e consequentemente necessitará de ser ajustado.

A engenharia de controlo do processo recorre normalmente a controladores por realimentação; os desvios das respostas em relação aos valores alvo estão normalmente autocorrelacionados e essa informação é normalmente utilizada para prever desvio futuros. A partir dos valores medidos das saídas autocorrelacionadas constroem-se modelos de séries temporais e estimam-se os respectivos parâmetros. Estes modelos são utilizados para obter previsões de erro quadrático mínimo de distúrbios futuros e o algoritmo de controlo é implementado de forma a cancelar os desvios previstos. Contudo, uma estratégia eficiente do processo de ajustamento deve ter em conta os aspectos económicos.

3.1.1.3 Modelo estocástico

Box e Jenkins propuseram um processo iterativo de três estágios para construção de um modelo a partir dos dados, identificação, estimação e diagnóstico de validação do modelo.

A identificação do modelo consiste em determinar as ordens *p,d,q* de um possível modelo autoregressivo integrativo média móvel ARIMA(*p,d,q*) a partir de funções de autocorrelação e autocorrelação parcial.

Uma vez identificadas as ordens do modelo, os respectivos parâmetros tem que ser estimados. Se a estimativa é feita com dados históricos, a estimativa diz-se off-line. Os parâmetros são estimados com recurso a ferramentas como estimativas dos mínimos quadrados e estimativas da máxima verosimilhança.

Se o modelo obtido é apropriado, os resíduos não devem conter qualquer informação. Se a autocorrelação foi completamente capturada, os resíduos deverão ser ruído branco. No diagnóstico de validação a função de autocorrelação dos resíduos é analisada e validada com recurso a testes estatísticos. Se os resíduos mostrarem autocorrelação significativa, então o modelo deve ser refeito e todos os três estágios devem ser repetidos até se encontrar um modelo satisfatório.

Modelo Integrativo Média Móvel ARIMA(0,1,1)
O modelo IMA(1,1) é um caso especial dos modelos ARIMA(*p,d,q*). Este modelo tem servido para representar uma larga gama de séries temporais utilizadas na prática, tais como procuras de mercado, preços, stocks, características de processos químicos e físicos (temperaturas, viscosidades, concentrações) etc. Este modelo tem-se mostrado também bastante apropriado para modelar distúrbios que ocorrem nos processos. Com o parâmetro AR igual a zero e os parâmetros I e MA iguais a um, o modelo pode ser representado pela forma

$$\nabla z_t = \varepsilon_t - \theta \varepsilon_{t-1}$$

O modelo é caracterizado por dois parâmetros θ e σ_ε. Este modelo pode igualmente ser reescrito numa forma mais conveniente

$$\nabla z_t = (1-\theta)\varepsilon_t$$

$$z_t = z_{t-1} + \varepsilon_t + -\theta \varepsilon_{t-1}$$

$$z_t = C^{te.} + \varepsilon_t + \lambda \sum_{j=1}^{t-1} \varepsilon_j$$

Intuitivamente, z_t é uma mistura de "*shocks"* aleatórios correntes com a soma de "*shocks"* aleatórios. Este modelo é bastante utilizado na previsão de distúrbios e caracterização da função de transferência de processos dinâmicos.

3.1.1.4 Aspectos económicos relacionados com EPC

O objectivo do EPC é ajustar o processo e manter as características da qualidade o mais próximo possível dos valores alvo. Na prática contudo, existem vários custos envolvidos. Esses custos deverão ser tidos em conta para se tomar uma decisão racional. Os principais custos envolvidos são os custos de não conformidade (*Off-target costs*), custos de ajustamento e custos de amostragem (Box & Luceño, 1997).

Os custos de não conformidade são custos que incorrem quando as características da qualidade se desviam dos valores pretendidos e são normalmente definidos de forma linear, proporcionalmente aos desvios em relação aos valores alvo.

O ajustamento do processo pode incorrer em custos significantes na vida real dos processos. Para além de custos de manipulação e operação dos actuadores, os ajustamentos podem requerer a paragem dos processos. Consequentemente, os ajustamentos frequentes podem ser uma situação a evitar. Em EPC, normalmente os

custos de ajustamento assumem-se fixos e independentes da magnitude do ajustamento.

Os custos de amostragem são aqueles que incorrem na obtenção dos valores numéricos finais das características da qualidade. Estes incluem os custos incorridos na amostragem do processo e do processamento de análises físicas ou químicas para se determinar leituras precisas e objectivas das medidas efectuadas. Quando os custos de amostragem são significativos, poderá ser indesejável uma frequência elevada de amostragem.

Estes custos variam fortemente de situação para situação. Enquanto que numa produção os custos de não conformidade podem ser dominantes, noutros processos, os custos de amostragem podem ser bastante elevados, enquanto noutros ainda, os custos de ajustamento podem ser bastante altos, envolvendo a paragem do processo ou custos relacionados com reparações.

3.1.1.5 Ajustamento por feedback com zona morta

Se não existirem custos de ajustamento nem custos de amostragem, será conveniente ajustar o processo em todos os períodos. Nesses casos, os controladores de variância mínima serão apropriados e eficientes mantendo o processo nos valores óptimos pretendidos. Em muitos casos práticos não é desejável ajustar o processo tão frequentemente devido aos custos envolvidos. Um controlador que apenas minimize as características da qualidade, negligenciando os outros custos, poderá não ter de grande utilidade prática.

Para gerir os custos globais do controlo por realimentação, foi então proposto um controlador com frequência de ajustamento limitada. No ajustamento por feedback condicionado, é colocada uma zona morta em redor dos valores objectivo. O processo apenas é ajustado se os desvios previstos caírem fora da zona morta. A espessura da zona morta é função dos custos de não conformidade do produto, dos custos de ajustamento e dos custos de amostragem.

3.1.1.6 Problema do ajustamento de setup

Nos processos produtivos é crucial a precisão do ajustamento das máquinas no inicio da produção de um lote. Um setup incorrecto pode resultar em severas consequências para o lote produzido. O efeito do erro de setup é induzir uma alteração na saída. Será necessário ajustar e corrigir, se possível, o processo que teve um erro de setup induzido no inicio do lote.

Considere-se um processo cujo setup foi ajustado antes da produção do lote e este setup está sujeito a erros. O erro de entrada originará um desvio em degrau na saída y_t do processo. O objectivo então é ajustar o processo para eliminar o offset induzido na saída. Suponha-se que se tem acesso à variável de controlo x_t e que esta tem um efeito direto na saída e que o efeito do controlo exercido é instantâneo. No ajustamento do setup, o objectivo é trazer rapidamente o processo para os valores pretendidos recorrendo à estimativa do offset com precisão. A amplitude do offset é estimada a partir dos dados observados. As observações estão sujeitas a variações inerentes ao processo e erros medidos. A precisão das estimativas do offset pode ser melhorada

com o incremento das observações disponíveis. A espera pela colecta de uma quantidade significativa de dados entra em conflito com o objectivo de trazer rapidamente o processo para os valores desejados. Uma estratégia óptima para esta situação passa por estimar sequencialmente o offset e ajustar o processo em conformidade.

Grubbs propôs uma elegante regra de ajustamento sequencial para resolver o problema de ajustamento do erro de setup, conhecida como regra harmónica de Grubb (Grubb, 1954/1983). A estratégia de ajustamento proposta consiste em ajustar o processo de acordo com a seguinte equação

$$x_{t+1} - x_t = \frac{-(y_t - T)}{t} \quad t = 1, 2, 3, \ldots$$

A regra de ajustamento implica que, após a produção da primeira parte, o processo é ajustado em função do desvio observado. Após a produção da segunda parte, o processo é ajustado em função de metade do desvio observado e assim por diante. O ajustamento segue a serie harmónica $\left[1, \frac{1}{2}, \frac{1}{3}, \ldots\right]$, sendo por isso conhecida como a serie harmónica de Grubb.

Foram feitos os seguintes pressupostos:

1. O processo é estável sem autocorrelação (sem dinâmica) ou alteração na média.
2. Os ajustamentos alteram a média do processo.
3. Os ajustamentos são exactos e implementados em todas as partes.

Os problemas de ajustamento e de setup tem tido recentes desenvolvimentos através de estudos de Pan e Del Castillo (Pan & Del Castillo, 2003) e Sulo e Vandevan (Sulo & Vandevan, 1999).

3.1.2 Controlo do processo lote a lote (run-to-run)

O controlo *run-to-run* é também uma forma discreta de controlo por realimentação no qual as acções de controlo são exercidas entre lotes com o objectivo de minimizar os desvios em relação aos valores alvo e a variabilidade dos processos. Este tipo de controlo tem tido as principais aplicações no contexto de produção de semicondutores. Neste contexto, tem-se verificado bastantes pesquisas envolvendo tipos de controlo de entradas e saídas múltiplas (sistema MIMO), (Moyne, del Castillo, & Hurwitz, 2001) embora a maioria das publicações se refiram a resultados simulados.

Controladores EWMA

Os controladores EWMA são os mais amplamente utilizados na indústria de produção de semicondutores. Estes controladores são simples e ainda bastante eficazes na manutenção do processo nos níveis desejados e redução da variabilidade. O procedimento no ajustamento do processo com controladores EWMA é o seguinte:

Considere-se um processo, que embora não linear, pode ser linearizado num determinado ponto de operação. Esse processo na vizinhança de ponto pode ser descrito pela seguinte equação:

$$\boldsymbol{y_t = \alpha + \beta x_{t-1} + \eta_t}$$

Onde $\boldsymbol{y_t}$ é o valor da característica da qualidade do processo (saída) para o lote t; $\boldsymbol{x_{t-1}}$ é a variável de controlo (entrada) determinada no fim do processamento do lote $t-1$; $\boldsymbol{\eta_t}$ é a perturbação do processo; $\boldsymbol{\alpha}$ é o offset do processo (ordenada na origem) e $\boldsymbol{\beta}$ é o declive ou ganho. Assume-se que $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$ são constantes ao longo do tempo. Estes valores são desconhecidos e tem que ser estimados a partir dos dados disponiveis.

O ganho do processo é similar a um coeficiente de regressão que relaciona a alteração da saída com a correspondente alteração na entrada. O ganho do processo é estimado através de desenho de experiências, análise de regressão e metodologia de resposta em superfície (Box & Draper, 2007).

Seja p_0 e b as estimativas iniciais de $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$. p_0 e b são tipicamente estimados pela metodologia dos mínimos quadrados a partir dos dados históricos. Tal como como em qualquer outro controlador nestas circunstâncias, a variável de controlo é ajustada para invalidar o desvio da resposta, ou seja

$$\boldsymbol{x_0 = \frac{T - p_0}{b}}$$

onde T é o valor de *Target*.

No controlador EWMA proposto, o parâmetro desconhecido $\boldsymbol{\alpha}$ é recursivamente estimado e actualizado e a variável de entrada é determinada ao fim de cada lote. A equação de estimação é a seguinte

$$\boldsymbol{p_t = w(y_t - b x_{t-1}) + (1 - w) p_{t-1}}$$

onde $0 \leq w \leq 1$ é o chamado factor de desconto.

O valor de offset estimado $\boldsymbol{p_t}$ é substituído na equação seguinte para determinar o valor da variável de controlo

$$\boldsymbol{x_t = \frac{T - p_t}{b}}$$

Como é notório, a ideia chave no controlador EWMA é que para um predeterminado ganho do processo, o offset e variáveis de entrada são actualizadas recursivamente. O valor esperado convergirá então assimptoticamente para o desejado valor de *target*.

Se as perturbações do processo seguirem o modelo não estacionário IMA(1,1)

$$\boldsymbol{\eta_t = \eta_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1}}$$

$$\boldsymbol{\varepsilon_t \sim N(0, \sigma_\varepsilon^2)}$$

e o viés do ganho estimado pode ser representado como $\xi = \beta/b$ sob a condição de que $0 < \beta/b < 2$, o ganho estimado está enviesado não mais que duas vezes o valor estimado, o factor óptimo de desconto é dado por

$$\boldsymbol{w_0 = \frac{b(1-\theta_1)}{\beta}}$$

Contudo, uma estimativa imprecisa dos parâmetros desconhecidos $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$ conduz a valores elevados do viés inicial

$$\boldsymbol{\frac{\alpha + \beta(T - p_0)}{\beta}}$$

e deverá demorar alguns lotes para o controlador EWMA levar o processo de volta ao *target*.

Controladores de duplo EWMA

Os processos de produção podem estar sujeitos a deriva determinística com o tempo e tender a afastar-se da referência. Estes fenómenos podem estar relacionados com o envelhecimento das máquinas ou deterioração das condições ideais de produção com o tempo. O objectivo do controlo por realimentação é ajustar as variáveis de controlo tal que a saída esteja tão próxima quanto possível da referência. A utilização de uma *single-EWMA* neste caso poderá não ser óptima porque não pode compensar a tendência determinística.

Considere-se um processo que está compensado e sujeito a distúrbios e deriva com o número de processamentos. O processo pode ser descrito pela seguinte expressão

$$\boldsymbol{y_t = \alpha + \beta x_{t-1} + \delta t + \eta_t}$$

Como definido anteriormente $\boldsymbol{y_t, \alpha, \beta, \eta_t, x_{t-1}}$ correspondem respectivamente à saída, ao *offset,* ao declive, aos distúrbios e à entrada. $\boldsymbol{\delta}$ é a taxa de deriva determinística. Os seguintes controladores de dupla EWMA aplicam-se a processos produtivos com deriva linear (Moyne, del Castillo, & Hurwitz, 2001)

$$\boldsymbol{x_t = \frac{T - p_t + D_t}{b}}$$

$$\boldsymbol{p_t = w_1(y_t - bx_{t-1}) + (1 - w_1)p_{t-1}} \qquad \boldsymbol{0 < w_1 < 1}$$

$$\boldsymbol{D_t = w_2(y_t - bx_{t-1} - p_{t-1}) + (1 - w_2)p_{t-1}} \qquad \boldsymbol{0 < w_2 < 1}$$

O controlador de duplo EWMA consiste num filtro para estimar a verdadeira saída do processo e num filtro para estimar a tendência. A previsão é igual à soma das duas componentes.

Duplo EWMA avançado

Em (Moyne, del Castillo, & Hurwitz, 2001), Ruey-Shan Guo, Argon Chen e Jing-Jung Chen apresentam o denominado "*Enhanced EWMA Controller*" que é um controlador baseado em filtros EWMA composto por dois módulos denominados por "*Dynamic-*

Tunning Loop module" e "*Run-by-Run Feedback Control module*". O modulo *Dynamic-Tunning Loop* utiliza primeiro uma carta de controlo EWMA para determinar a existência de alterações médias ou grandes no processo. Se existir uma alteração média ou grande no processo, e o erro é controlável, então parâmetro de controlo (factor de desconto w) é alterado para valores altos e inicia o módulo *Dynamic-Tunning Loop* de modo a trazer o processo rapidamente aos valores de referência. No modulo *Run-by-Run Feedback Control* o parâmetro de controlo w é sintonizado baseado no estado corrente do *Dynamic-Tunning Loop.* Durante a execução do *Dynamic-Tunning Loop,* o parâmetro de controlo w vai diminuindo até atingir o valor predeterminado w_0 a partir do qual *Run-by-Run Feedback* volta às condições iniciais. Se o erro não é controlável, o processo pára disparando assim uma acção de controlo externa.

Controlo Run-to-Run para pequenas produções
Nos controladores de base EWMA apresentados até aqui pressupunha-se que os processos atingissem o estado estacionário com um apropriado factor de desconto, mas para atingir o estado estacionário poderá ter que se processar alguns lotes até que o processo atinja os valores de referência. Este tempo de resposta poderá ter consequências severas nas pequenas produções, frequentemente encontradas até na indústria de semicondutores, onde esta matéria tem tido os desenvolvimentos mais recentes. O objectivo dos controladores IIIA (Internal Intercept Iteratively Adjusted) e controladores de EWMA variável é superar esta negligência e reduzir a alta taxa de retrabalho nas produções iniciais (Krishna B. Misra, 2008).

3.1.2.1 EPC e SPC como ferramentas complementares

O SPC e o EPC podem trazer grandes benefícios quando utilizados como ferramentas complementares. Com esta combinação consegue-se verificar a adequação dos ajustamentos e simultaneamente identificar causas atribuídas de alterações no desempenho.

Existem críticas argumentando que a implementação do EPC impede de detectar oportunidades de melhorar os processos pela identificação de causas especiais e consequente remoção. O EPC tem sido descrito como uma banda par cobrir a ferida e não para a curar. A implementação de EPC e controlo por *feedback* pode cancelar a verdadeira natureza dos distúrbios que afectam o processo. Essa situação pode ser evitada pela monitorização das variáveis de controlo e dos desvios das respostas em relação às referências. Qualquer ponto fora de controlo nas variáveis de controlo poderá corresponder a uma compensação excessiva para uma perturbação que poderá desencadear um alarme correspondente a uma causa especial no sistema. De modo similar, qualquer erro fora do normal será visto através das cartas de monitorização dos desvios em relação às referências. Outra crítica é que a implementação do EPC resultará numa compensação sub-óptima. Isso verificar-se-á se os modelos dos distúrbios estão errados ou os parâmetros estimados sem precisão. Na maior parte dos casos, assume-se que os distúrbios são modelados por um modelo IMA e então utiliza-se uma EWMA para prever os distúrbios. Normalmente a constante de amortecimento é inferior a 0.3, consequentemente a inflação da variância do processo devido aos erros de modelação não será severa. Adicionalmente, o modelo EWMA fornece uma boa estimativa mesmo que o modelo não seja precisamente um IMA (Krishna B. Misra, 2008).

A conveniência de utilização de EPC e SPC num processo depende dos seguintes factores:

1. Quando os custos de ajustamento e/ou erros de ajustamento são elevados, é melhor não implementar o EPC.
2. Quando os erros medidos são elevados não é aconselhável utilizar o EPC.
3. Quando a amostragem é lenta relativamente à dinâmica do processo, o processo observado não exibe autocorrelação e controlo estatístico será suficiente.

De notar que a redução da frequência de amostragem para propósitos de SPC será útil apenas se o processo original é estacionário. A redução da frequência de amostragem de um processo não estacionário ainda produz processos não estacionários. Uma estratégia de ajustamento baseada no EPC será aconselhável em tais casos (Krishna B. Misra, 2008).

As cartas de controlo podem ter um papel efectivo na redução da variação do processo quando existem causas especiais em processos controlados por EPC. O denominado ASPC (*Algoritmic Statistical Process Control*) é uma estrutura que unifica o SPC e o EPC. É um sistema que utiliza o EPC para regular o processo e SPC para monitorizar o processo controlado com a finalidade de detectar qualquer alteração do processo em relação ao modelo assumido e revê-lo se necessário. Existem pelo menos dois métodos utilizados na integração do SPC com EPC, o primeiro consiste em monitorizar a saída do processo controlado por EPC, a segunda consiste em monitorizar as acções de controlo do EPC. Um dos problemas com a monitorização das respostas é que quando existem causas especiais, as respostas estão contaminadas pelas acções de controlo, o que resultará numa janela mais pequena para as oportunidades de detecção. No entanto as saídas estão correlacionadas com as entradas, pelo que através da monitorização das acções de controlo, estas situações podem perfeitamente ser detectadas.

A performance da monitorização de um processo controlado por EPC pode depender da monitorização das respostas ou acções de controlo, da estratégia de controlo empregue (MMSE ou PID) e da estrutura de autocorrelação subjacente.

SPC Algoritmico

(Vander Weil, 1992) propôs um algoritmo de controlo baseado no MMSE (*Minimum Mean Square Error*) e um esquema de monitorização da saída por uma carta CUSUM, a que se chamou *Algorithmic* SPC ou ASPC. (Montgomery D. C., 1994) Investigou dois tipos de causas especiais, uma alteração contínua e uma tendência linear, e demonstrou que os procedimentos combinados melhoram a performance em relação ao procedimento apenas do EPC (ver também: (MacGregor, 1987), (MacGregor J. F., 1990), (Box G. E., 1992), (Capilla, Ferrer, Romero, & Hualda, 1999), (Montgomery D. C., 2001), (Chen, 2002)). Adicionalmente, a monitorização de séries autocorrelacionadas tem semelhança visível com o ajustamento do processo. A performance das várias cartas de monitorização tem sido investigada para casos especiais de nível de alteração ((Alwan, 1988), (Wardell, 1994), (Vander Weil S. , 1996), (Atienza, 1998)

3.2 Termos e definições

3.2.1 Notação

AAI	Intervalo médio entre ajustes	*Average Ajustment Interval*
AIC	Critério de Informação de Akaike	*Akaike Information Criterion*
ARIMA(p,d,q)	Modelo autorregressivo de ordem p, diferenciação de ordem d e de médias móveis de ordem q	*Autorregressive Moving Average Model*
ARL	Número médio ao fim do qual se detecta	*uma situação fora de controlo Average Run Length*
ARMAX	Modelo autorregressivo, de médias móveis com variável Exógena	*Autorregressive Moving Average Model with Exogene Variable*
ARX	Modelo autorregressivo com variável Exógena	*Autorregressive Model with Exogene Variable*
AR(p)	Modelo autorregressivo de ordem p	*Autorregressive Model*
BJ	Modelo de Box e Jenkins	*Box Jenkins Model*
Carta CUSUM	Carta de Somas Acumuladas	*Cumulative Sums Chart*
Carta EWMA	Carta de Média Móvel Exponencialmente Amortecida	*Exponential Weight Moving Range Chart*
EPC	Engenharia de controlo de processo	*Engeneering Process Control*
FAC	Função de Autocorrelação	*ACF – Autocorrelation Function*
FACP	Função de Autocorrelação Parcial	*PACF - Partial Autocorrelation*
Function FCC	Função de Correlacção-Cruzada	*CCF – Cross-Correlation Function*
fdp	Função densidade de probabilidade	
FT-R	Função de Transferência-Ruído	
LC	Linha Central	*CL – Center line*
LIC	Limite Inferior de Controlo	*LCL –Lower Control Limit*
LSC	Limite Superior de Controlo	*UCL –Upper Control Limit*
ISD	Percentagem de inflação no desvio padrão	*Increase of the Standard Deviation*
MSE	Erro médio quadrático	*Mean Squared Error*
MIMO	Várias variáveis de saída e várias variáveis de entrada	*Multiple Input Multiple Output*
MISO	Uma variável de saída e várias variáveis de entrada	*Multiple Input Single Output*
MR	Amplitudes Móveis	*Moving Range*
PCA	Análise em Componentes Principais	*Principal Component Analysis*

SISO	Uma variável de saída e uma variáveis de entrada	*Single Input Single Output*
SPC	Engenharia de controlo do processo	*Statistical Process Control*
SPE	Erro de previsão quadrático	*Squared Prediction Error*
VM	Variância Mínima	

3.2.2 Conceitos e definições

Processo estacionário
Um processo é estacionário se os seus primeiro e segundo momentos são invariantes no tempo. Por outras palavras, um processo estocástico $y(t)$ é estacionário se se verificar simultaneamente

$$\begin{cases} E(y_t) = \mu & \text{Para todo } \boldsymbol{t} \\ E[(y_t - \mu)(y_{t-h} - \mu)'] = \Gamma_y(h) = \Gamma_y(-h)' & \text{Para todo } \boldsymbol{t} \text{ e } \boldsymbol{h} = \mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots \end{cases}$$

A primeira condição significa que todo o y_t tem o mesmo vector média μ e a segunda condição requer que a autocovariância do processo não dependa de t mas apenas do período de tempo h de que os dois vectores $\boldsymbol{y_t}$ e $\boldsymbol{y_{t-h}}$ estão separados (Lütkepohl, 2007).

Autocovariância e Autocorrelação
O prefixo auto significa uma acção reflectiva nele próprio, assim a autocovariância é a covariância que o processo tem com ele próprio. A função autocovariância, denominada por $(\gamma_{t,k})$ dá a medida da dependência linear (associação linear) entre duas variáveis do mesmo processo Y_t e Y_{t+k}, que estão separadas por k períodos ou ***lags***, como função de k. para qualquer processo discreto, define-se

$$\boldsymbol{\gamma_{t,k} = Cov[Y_t, Y_{t+k}] = E[(Y_t - E[Y_t])(Y_{t+k} - E[Y_t])]} \qquad k = 0, \pm 1, \pm 2 \ldots$$

Se o processo é estritamente estacionário, a média μ é constante e a autocovariância reduz-se a

$$\boldsymbol{\gamma_{t,k} = Cov[Y_t, Y_{t+k}] = E[(Y_t - \mu)(Y_{t+k} - \mu)]} \qquad k = 0, \pm 1, \pm 2 \ldots$$

Pelo que a função autocovariância depende apenas da *lag* k, sendo que, para $k = 0$ tem-se $\boldsymbol{\gamma_0 = E\left[(Y_t - \mu)^2\right] = Var(Y_t) = \sigma^2}$.

Normalmente torna-se mais simples trabalhar com a autocovariância normalizada que se obtém dividindo a autocovariância do processo pela própria variância do processo, neste caso $\boldsymbol{\gamma_0}$, obtendo-se assim a função denominada autocorrelação. Para um processo estacionário define-se a função autocorrelação ρ dada por

$$\boldsymbol{\rho_k = \frac{\gamma_k}{\gamma_0}} \qquad k = 0, \pm 1, \pm 2 \ldots$$

Onde $-1 \leq \boldsymbol{\rho_k} \leq 1$. Dado que $\boldsymbol{\gamma_k = \gamma_{-k}}$ o que implica que $\boldsymbol{\rho_k = \rho_{-k}}$, estas funções normalmente estão definidas apenas para *lags* positivas. Tanto a autocorrelação como a autocovariância na *lag* k dão o grau de associação linear entre duas variáveis do mesmo processo, Y_t e Y_{t-k}, separadas que estão separadas k períodos.

Nas relações entre variáveis aleatórias é conveniente relembrar as seguintes implicações:

1. Se Y_t e Y_{t-k} são independente então são não correlacionadas, isto é, $\boldsymbol{\rho_k = 0}$ qualquer que seja k.
2. Se Y_t e Y_{t-k} são correlacionados isto é, $\boldsymbol{\rho_k \neq 0}$ para algum k, então Y_t e Y_{t-k} são dependentes.

O desvio padrão da média de um processo $(\bar{Y})$ é fortemente afectado pela autocorrelação. (Bartlett, 1946) mostrou que

$$\boldsymbol{\sigma_{\bar{Y}} = \sqrt{\frac{\gamma_0}{N}\left[1 + 2\sum_{k=1}^{N}\left(1 - \frac{k}{N}\right)\rho_k\right]}}$$

Para grandes amostragens, $N \gg k \Rightarrow \boldsymbol{k/N \simeq 0}$, pode-se fazer a seguinte aproximação:

$$\boldsymbol{\sigma_{\bar{Y}} = \sqrt{\frac{\gamma_0}{N}\left[1 + 2\sum_{k=1}^{\infty}\rho_k\right]}} \qquad \textbf{(3.1)}$$

Deste modo, se um processo é completamente não correlacionado, ou seja, $\boldsymbol{\rho_k = 0}$ para todo o k, o desvio padrão da média do processo será dado pelo usual $\rho/\sqrt{N}$.

A estimativa da função autocovariância pode ser obtida através da função autocovariância amostral c_k dada por (Del Castillo, 2002)

$$\boldsymbol{c_k = \hat{\gamma}_k = \frac{1}{N}\sum_{t=1}^{N-k}(y_t - \bar{Y}) \cdot (y_{t+k} - \bar{Y})} \qquad \boldsymbol{k = 0, 1, 2, \ldots} \qquad \textbf{(3.2)}$$

De forma similar, a função de autocorrelação é dada por

$$\boldsymbol{r_k = \hat{\rho}_k = \frac{c_k}{c_0}} \qquad \boldsymbol{k = 0, 1, 2, \ldots}$$

Ruído branco vectorial (Multivariado)

O processo vectorial ruído branco é definido como uma sequência de vectores aleatórios independentes, $\ldots, a_1, \ldots, a_t, \ldots$ em que $a_t = (a_{1t}, \ldots, a_{kt})'$, tal que

$E[a_t] = 0,\ E[a_t a_t^{'}] = \Sigma$, onde Σ é a matriz de covariância kxk assumida positiva definida, e $E[a_t a_{t+l}^{'}] = 0$ para $l \neq 0$ devido à independência. Consequentemente as suas matrizes de covariância são dadas por

$$\boldsymbol{\Gamma(l) = E[a_t a_{t+l}^{'}] = \begin{cases} \Sigma & se\ \ l = 0 \\ 0 & se\ \ l \neq 0 \end{cases}}$$

Operador "vec" e produto de *Kroneker* de Matrizes
Define-se o operador "*vec*" que transforma uma matriz num vector coluna colocando (ou empilhando) as colunas da matriz uma por baixo das outras. Ou seja, se B é uma matriz de dimensão $m \times k$ definida por $B = (b_1, \dots, b_k)$ onde b_i é a *i*-ésima coluna de B, define-se o vector coluna de dimensão $mk \times 1$ dado pela expressão

$$\boldsymbol{b = vec(b) = (b_1^{'}, \dots, b_k^{'})^{'}} \quad \textbf{(3.3)}$$

O produto de *Kroneker* de duas matrizes é definido como se segue. Seja $A = (a_{ij})$ uma matriz de dimensão $m \times n$ e $C = (c_{ij})$ uma matriz de dimensão $p \times q$. Então o produto de *Kroneker* de *A* e C, cuja notação é $A \otimes C$, é a matriz $mp \times nq$ da forma

$$\boldsymbol{A \otimes C = [(a_{ij} C)]} \quad \textbf{(3.4)}$$

Uma propriedade bastante útil relacionada com o operador "*vec*" é que se Z=ABC, então

$$\boldsymbol{vec(Z) = vec(ABC) = (C^{'} \otimes A) vec(B)} \quad \textbf{(3.5)}$$

De notar ainda que $\boldsymbol{tr(AB) = [vec(A^{'})]^{'} vec(B) = [vec(B^{'})]^{'} vec(A)}$, consequentemente, utilizando (3.5), tem-se que $tr(ABCB^{'}) = \boldsymbol{[vec(B^{'})]^{'} (C^{'} \otimes A) vec(B)}$.

3.3 Estrutura

Em linhas gerais, a presente dissertação é constituída por quatro partes mais Anexos. A primeira parte, de reduzida dimensão e de natureza predominantemente informativa, inclui os índices, o enquadramento do tema do trabalho, uma introdução ao tema e os trabalhos mais relevantes nesta área e ainda algumas abreviaturas, termos e definições utilizadas ao longo este trabalho.

A segunda parte que constitui a estrutura principal deste trabalho, é constituída pelos capítulos 3 a 8, e está dividida em dois grandes blocos. O bloco 1, que vai do capítulo 4 ao capítulo 7, apresenta os fundamentos teóricos onde assenta toda a parte experimental e investigação realizada. O bloco 2 é composto pelo capítulo 8 e descreve o trabalho efectuado em termos práticos.

O bloco 1 por sua vez está dividido em 3 sub-blocos. O sub-bloco 1 é composto pelos capítulos 4 e 5 e aborda temas relativos a modelos matemáticos de processos e sistemas de controlo no domínio escalar, ou seja, trata-se de modelos que apenas usam uma variável, ou, quanto muito trata-se de modelação de sistemas com uma variável de entrada e uma variável de saída. No sub-bloco 2 faz-se a análise do mesmo tipo de

sistemas mas extrapolando para o plano multivariado. No sub-bloco 3 apresenta-se várias metodologias de abordagem ao tema objectivo deste trabalho que é a integração do controlo estatístico com engenharia de controlo.

O bloco 2 descreve-se e caracteriza-se o processo que serviu de base ao estudo efectuado, apresenta-se os desenvolvimentos práticos efectuados e as conclusões preliminares.

Bloco 1
O bloco 1 é composto pelos capítulos 4, 5, 6 e 7. No capitulo 4 introduz-se o conceito de modelação de sistemas dinâmicos, faz-se uma abordagem à transformada de Laplace e á modelação de sistemas dinâmicos no espaço das fazes (domínio s). A partir desta abordagem passa-se para a modelação discreta com a introdução da transformada z e faz-se a ponte para a modelação no domínio do tempo introduzindo-se os modelos de séries temporais.

O subcapítulo 4.2 apresenta os principais conceitos subjacentes à teoria das séries temporais e as metodologias de identificação, estimação de parâmetros, teste e validação de modelos autoregressivos (AR), média móvel (MA), e modelos mistos ARMA e ARIMA.

O subcapítulo 4.3 introduz o conceito de função de transferência e a sua relação com a resposta ao impulso. Introduz os modelos de variável exógena ARX e ARMAX e a metodologia de identificação de funções de transferência de sistemas dinâmicos. O ponto 4.3.3 apresenta metodologias de estimação recursiva (on-line) de modelos ARMA e ARMAX. Embora esta metodologia se situe na secção referente a sistemas univariáveis, ela é também aplicável a sistemas multivariávais.

O capítulo 5 aborda um dos problemas principais deste trabalho: como ajustar um processo com a garantia de que os ajustamentos efectuados são os mais indicados (óptimos). Este capítulo além de introduzir o conceito de controlo por realimentação (*feedback*), faz a abordagem a algumas das metodologias de controlo aplicáveis aos modelos estudados até este ponto como sejam os controladores teóricos de variância mínima ou o modelo de controlo de Clarke e Gawthrop.

O sub-bloco 2 do bloco 1 é composto exclusivamente pelo capítulo 6 e é talvez o capítulo core deste trabalho. A secção 6.1 começa por apresentar as séries temporais multivariadas, aborda-se algumas das dificuldades inerentes à identificação e representação de forma única dos modelos mistos ARMA e estabelece-se a metodologia para a construção dos modelos preliminares passando depois à estimação dos modelos finais com recurso a estimativas da máxima verosimilhança. A exemplo do que foi apresentado para o caso univariado, a secção 6.2 aborda a introdução das variáveis exógenas nos modelos VARMA (uma das designações aplicadas aos modelos ARMA multivariados - Vector ARMA) passando a denominar-se por modelos VARMAX. Esta secção faz ainda uma abordagem aos sistemas de controlo aplicáveis este tipo de processos e finaliza com a apresentação de técnicas de especificação e validação de modelos VARMAX.

O sub-bloco 3 do bloco 1 é composto exclusivamente pelo capítulo 7. Este capítulo apresenta algumas metodologias básicas de integração do controlo estatístico com

engenharia de controlo. O subcapítulo 7.1 faz uma abordagem aos custos envolvidos nas estratégias de controlo e apresenta soluções para determinar os parâmetros de controlo que minimizam os custos envolvidos. No subcapítulo 7.2 faz-se uma ligeira abordagem às cartas de monitorização Cuscore com ênfase na sua vocação para detectar alterações nos parâmetros dos modelos de séries temporais ARMA. No subcapítulo 7.3 apresenta-se o conceito de controlo adaptativo que será uma das áreas onde a integração EPC/SPC pode ter particularmente sucesso. Nessa mesma secção apresenta-se ainda o caso mais estudado nesta área nos últimos anos (Del Castillo & Yeh, An adaptive run-to-run optimizing controller for linear and nonlinear semiconductor processes, 1998) e que constitui uma das maiores referências deste trabalho.

Bloco 2
Neste bloco apresenta-se alguns resultados do trabalho prático efectuado. A base do trabalho prático efectuado assenta basicamente nos fundamentos teóricos apresentados na secção 6. No ponto 8.1 faz-se uma análise do processo de estudo e estabelecem-se as indicações preliminares sobre a estrutura do modelo. Nos pontos 8.2 e 8.3 procede-se à modelação e validação do modelo que será utilizado como o modelo representativo do processo nas simulações a efectuar. No ponto 8.4 apresentam-se os resultados e conclusões preliminares resultantes de algumas das simulações efectuadas.

A terceira parte é composta pelo capítulo 9, designado como “Conclusões, Recomendações e Trabalho Futuro”. Nesta parte apresenta as conclusões gerais do trabalho realizado e das limitações verificadas. Apresenta-se ainda algumas situações que se pretendia abordar mas que não foi possível devido às condicionantes existentes. Finalmente especula-se um pouco sobre o que se poderá fazer em sequência deste trabalho e respectivas potencialidades.

A quarta parte é composta pelas referências bibliográficas.

4 Modelos Matemáticos de Processos

4.1 Introdução

Um modelo matemático de um sistema dinâmico pode ser definido como um conjunto de equações que represente com precisão a dinâmica do sistema, ou no mínimo, com relativa precisão. Um modelo não é único para um determinado sistema. Um sistema pode ser representado nas mais diversas formas e, portanto, em diversos modelos matemáticos, dependendo da perspectiva.

É possível melhorar a precisão de um modelo matemático incrementando a sua complexidade, incluindo, em alguns casos, centenas de equações para descrever um sistema completo. Na construção de um modelo matemático, há que ter o compromisso entre a simplicidade do modelo e a precisão dos resultados da análise. Se não for necessário uma precisão extrema é preferível a obtenção de um modelo matemático que se adeqúe ao problema em consideração.

Sistemas lineares

Um sistema é considerado **sistema linear** se se aplica o princípio da sobreposição. O princípio da sobreposição estabelece que a resposta produzida pela utilização simultânea de duas diferentes funções é a soma das duas respostas individuais.

Uma equação diferencial é linear se os coeficientes são constantes ou funções apenas das variáveis independentes. Sistemas dinâmicos compostos por componentes lineares invariantes no tempo podem ser descritos por equações diferencias invariantes no tempo (coeficientes lineares constantes). A esses sistemas atribui-se o nome de **sistemas lineares invariantes no tempo**. Aos sistemas dinâmicos representados por equações diferenciais cujos coeficientes são funções do tempo atribui-se o nome de **sistemas lineares variantes no tempo**, por exemplo um sistema em movimento em que a massa total vai variando com o tempo devido ao consumo de combustível.

Um sistema é não linear se o princípio da sobreposição não se aplica. Para um sistema não linear, a resposta de duas entradas não podem ser calculadas pelo tratamento das entradas individuais e somando o resultado.

Em engenharia de controlo, a operação normal de um sistema não linear pode ser à volta de um ponto de equilíbrio, e os sinais podem ser considerados pequenos sinais à volta do equilíbrio. Nestes casos pode ser possível o sistema não linear num sistema linear. Nesses caos, o sistema linear é equivalente ao sistema não linear dentro dos limites da região de operação.

4.1.1 Sistemas de controlo contínuo

A dinâmica dos diversos sistemas, sejam mecânicos, eléctricos, térmicos, económicos, biológicos, etc. pode ser descrita em termos de **equações diferenciais** que poderão ser derivadas a partir das leis da física que descrevem determinado sistema, por exemplo, leis de Newton para sistemas mecânicos e leis de Kirchhoff's para sistemas eléctricos. Há, no

entanto, que ter sempre em mente que a obtenção de um razoável modelo matemático é a parte mais importante de toda a análise.

Os modelos matemáticos podem assumir as mais diversas formas dependendo da particularidade do sistema e das circunstancias, um determinado modelo pode ser mais apropriado que outro.

4.1.1.1 Transformada de Laplace

A transformada de Laplace é um método operacional com imensas vantagens na solução de equações diferenciais lineares. Através da transformada de Laplace pode-se converter qualquer função comum, tais como funções sinusoidais, funções exponenciais, etc. em funções algébricas de uma variável complexa s**.** Operações como diferenciação e integração, podem substituídas por operações algébricas no plano complexo. Deste modo, uma equação diferencial linear pode ser transformada numa função algébrica de variável complexa s. Em sentido contrario, se uma equação algébrica para a variável dependente, a solução da equação diferencial correspondente (a transformada inversa de Laplace da variável dependente) pode ser encontrada com recurso a uma tabela de transformadas de Laplace ou com recurso a técnicas de expansão de fracções parciais (Ogata, 1997).

Uma vantagem do método da transformada de Laplace é o uso de técnicas gráficas para a previsão do desempenho do sistema sem a necessidade de obtenção da solução das equações diferenciais. Outra das vantagens é que, uma vez resolvida a equação diferencial, tanto a componente transiente como a componente estacionária da solução podem ser obtidas simultaneamente.

Definindo-se:

f(t)	Uma função do tempo *t* tal que *f(t) = 0* para *t < 0*
s	Uma variável complexa
$\mathcal{L}$	Um símbolo operacional que indica que indica que a quantidade que lhe sucede é para ser transformada pelo integral de Laplace $\int_0^\infty e^{-st}dt$
F(s)	A transformada de Laplace de *f(t)*

A transformada de Laplace de *f(t)* será dada por:

$$\mathcal{L}[f(t)] = F(s) = \int_0^\infty e^{-st}dt\ [fTt)] = \int_0^\infty f(t)\, e^{-st}dt$$

Ao processo inverso de determinar a função no domínio do tempo *f(t)* partir da transformada de Laplace dá-se o nome de "transformada inversa de Laplace", é indicado pelo símbolo operacional $\mathcal{L}^{-1}$ e pode ser determinado pelo seguinte integral de inversão:

$$\mathcal{L}^{-1}[F(s)] = f(t) = \frac{1}{2\pi j}\int_{c-j\infty}^{c+j\infty} F(s)e^{st}ds, \qquad para\ t > 0$$

A transformada de Laplace de uma função $f(t)$ existe se o integral de Laplace converge. O integral convergirá se $f(t)$ é seccionalmente continua num intervalo finito no domínio $t > 0$ e se é de ordem exponencial quando t tende para infinito. Uma função $f(t)$ diz-se ser de ordem exponencial se uma constante real, positiva σ existe tal que a função

$$e^{-\sigma t}|f(t)|$$

Tende para zero quando t tende para infinito.

Função degrau
Considere-se a seguinte função degrau:

$f(t) = 0$ *para* $t < 0$

$f(t) = A$ *para* $t > 0$

onde *A* é uma constante. De notar que este caso é um caso especial da função $At^{-\alpha t}$, onde $\alpha = 0$. A função degrau é definida em $t = 0$. A respectiva transformada de Laplace é dada por

$$\mathcal{L}[A] = \int_0^\infty Ae^{-st}dt = \frac{A}{s}$$

No desenvolvimento deste integral, assume-se que a parte real de *s* é maior que zero o que implica que $lim_{t\to\infty} e^{-st}$ é zero. Esta função é válida em todo o plano *s* excepto no pólo $s = 0$.

À função degrau com amplitude unitária (*A=1*) dá-se o nome de degrau unitário. A função degrau unitário que ocorre no ponto $t = t_0$ é frequentemente descrita como $1(t - t_0)$. Seguindo a mesma notação, uma função degrau de amplitude *A* que ocorre em $t = 0$ pode ser descrita como $f(t) = A1(t)$. A transformada de Laplace da função degrau unitário, que nesta notação poderá ser definida como

$1(t) = 0$ *para* $t < 0$

$1(t) = 1$ *para* $t > 0$

é $1/s$ ou

$$\mathcal{L}[1(t)] = \frac{1}{s}$$

Fisicamente, uma função degrau que ocorre em $t = 0$ corresponde a um sinal constante aplicado ao sistema instantaneamente no instante *t* igual a zero.

Função transladada
Especial importância tem a função transladada dada por $f(t - \alpha)1(t - \alpha)$, onde $\alpha \geq 0$. Esta função é zero para $t < \alpha$.

Por definição, a transformada de Laplace de $f(t - \alpha)1(t - \alpha)$ é

$$\mathcal{L}[f(t-\alpha)1(t-\alpha)] = \int_0^\infty f(t-\alpha)1(t-\alpha)e^{-st}dt$$

Fazendo a mudança da variável independente *t* para τ, onde $\tau = t - \alpha$, obtém-se

$$\int_0^\infty f(t-\alpha)1(t-\alpha)e^{-st}dt = \int_{-\alpha}^\infty f(\tau)1(\tau)e^{-s(\tau+\alpha)}d\tau$$

Assumindo-se que $f(t) = 0$ para $t < 0$, $f(\tau)1(\tau) = 0$ para $\tau < 0$. Então pode-se mudar o limite inferior de integração de $-\alpha$ para 0

$$\begin{aligned}\int_{-\alpha}^\infty f(\tau)1(\tau)e^{-s(\tau+\alpha)}d\tau &= \int_0^\infty f(\tau)1(\tau)e^{-s(\tau+\alpha)}d\tau \\ &= \int_0^\infty f(\tau)e^{-s\tau}e^{-\alpha s}d\tau \\ &= \boldsymbol{e^{-\alpha s}\int_0^\infty f(\tau)e^{-s\tau}d\tau = e^{-\alpha s}F(s)}\end{aligned}$$

onde

$$F(s) = \mathcal{L}[f(t)] = \int_0^\infty f(t)e^{-st}dt$$

Concluindo-se que

$$\mathcal{L}[f(t-\alpha)1(t-\alpha)] = e^{-\alpha s}F(s), \qquad \alpha \geq 0$$

Esta equação estabelece que translação da função de tempo $f(t)1(t)$ por α (onde $\alpha \geq 0$) corresponde à multiplicação da transformada de $F(s)$ por $e^{-\alpha s}$.

Função pulso

Considere-se a função pulso

$$f(t) = \frac{A}{t_0}, \qquad para \;\; 0 < t < t_0$$
$$f(t) = 0, \qquad para \;\; t < 0, t_0 < t$$

Onde A e t_0 são constantes.

A função pulso aqui descrita deve ser vista como uma função degrau de amplitude A/t_0 que começa em $t = 0$, e que é sobreposta por outra função degrau negativa, com a mesma amplitude em $t = t_0$, isto é

$$f(t) = \frac{A}{t_0}1(t) - \frac{A}{t_0}1(t-t_0)$$

A respectiva transformada de Laplace será dada por

$$\begin{aligned}\mathcal{L}[1(t)] &= \mathcal{L}\left[\frac{A}{t_0}1(t)\right] - \mathcal{L}\left[\frac{A}{t_0}1(t-t_0)\right] \\ &= \frac{A}{t_0 s} - \frac{A}{t_0 s}e^{-st_0} \\ &= \frac{A}{t_0 s}(1-e^{-st_0}) \end{aligned} \quad \textbf{(4.1)}$$

Função impulso

A função impulso é um caso limite especial da função pulso. Considere-se a função:

$$\begin{aligned} g(t) &= \lim_{t_0 \to 0} \frac{A}{t_0}, \qquad para \;\; 0 < t < t_0 \\ &= 0, \qquad para \;\; t < 0, \qquad t_0 < t \end{aligned}$$

Se a amplitude da função impulso for A/t_0 e a duração t_0, a área abaixo do impulso será igual a A. Como a duração t_0 tende para zero, a altura for A/t_0 tende para infinito, mas a área mantém-se igual a A.

A transformada de Laplace da Equação (4.1) para esta função impulso será dada por:

$$\begin{aligned}\mathcal{L}[g(t)] &= \lim_{t_0 \to 0}\left[\frac{A}{t_0}(1-e^{-st_0})\right] \\ &= \lim_{t_0 \to 0}\frac{\frac{d}{dt_0}[A(1-e^{-st_0})]}{\frac{d}{dt_0}(t_0 s)} = \frac{As}{s} = A\end{aligned}$$

Assim, a transformada de Laplace da função impulso será igual à área subjacente à função.

À função impulso cuja área é igual a um, dá-se o nome “função impulso unitária” ou “função delta de Dirac”. A função impulso unitária que ocorre em $t = t_0$ denomina-se por $\delta(t - t_0)$, e satisfaz as seguintes condições:

$$\delta(t-t_0) = 0, \qquad para\; t \neq t_0$$

$$\delta(t-t_0) = \infty, \qquad para\; t = t_0$$

$$\int_{-\infty}^{\infty} \delta(t-t_0)dt = 1$$

De salientar que tal função com amplitude infinita e duração zero é ficção matemática que não ocorre em sistemas físicos. Se, contudo, tivermos um pulso na entrada de sistema com uma amplitude bastante grande, com uma duração bastante pequena comparada com a constante de tempo do sistema, então pode-se aproximar o pulso de entrada a uma função impulso.

O conceito de impulso é bastante útil na diferenciação de funções descontínuas, podendo inclusive considerar-se a função impulso como a derivada da função degrau unitário *1(t - t_0)* no ponto de descontinuidade *t* = t_0 ou seja

$$\delta(t - t_0) = \frac{d}{dt} 1(t - t_0)$$

Inversamente, se a função impulso unitário $\delta(t - t_0)$ é integrada, o resultado será a função degrau unitário $1(t - t_0)$.

Teorema da diferenciação real
A transformada de Laplace da derivada duma função *f(t)* é dada por:

$$\mathcal{L}\left[\frac{d}{dt} f(t)\right] = sF(s) - f(0) \quad \textbf{(4.2)}$$

Onde *f(0*) é o valor de *f(t)* em *t = 0*. Se todos os valores iniciais e as suas derivadas são iguais a zero, então a transformada de Laplace da *n*-ésima derivada de *f(t)* será dada por $s^n F(s)$.

As demonstrações deste teorema e seguintes podem ser consultadas em (Ogata, 1997) entre outros.

Teorema do valor final
Se *f(t)* e *df(t)/dt* são "Laplace transformáveis", se *F(s)* é a transformada de Laplace de *f(t)* e se $\lim_{t\to\infty} f(t)$ existe, então:

$$\lim_{t\to\infty} f(t) = \lim_{s\to 0} sF(s)$$

Este teorema relaciona o comportamento no estado estacionário de *f(t)* com o comportamento de *sF(s)* na vizinhança de *s = 0*. Este teorema aplica-se apenas se e só se $\lim_{t\to\infty} f(t)$ existe. Se todos os pólos de *sF(s)* se situam no lado esquerdo (real(s) menor que zero) do plano complexo, o $\lim_{t\to\infty} f(t)$ existe. Se *sF(s)* tiver no eixo imaginário ou à direita do eixo imaginário, *sF(s)* exibirá, respectivamente, comportamento oscilatório ou exponencial no domínio do tempo *(t)*, e o $\lim_{t\to\infty} f(t)$ não existe. O teorema do valor final não se aplicará em tais casos. Suponha-se que f(t) é uma função sinusoidal, *sin(ωt)*, a função *sF(s)* terá pólos em *s* = ±*j*ω e $\lim_{t\to\infty} f(t)$ não existe. Neste caso, este teorema não se aplica.

Teorema do valor inicial
Este teorema é a contraparte do teorema do valor final. Através deste teorema pode-se obter o valor de *f(t)* em *t = 0*+. Este teorema não nos dá exactamente o valor em *t = 0*, mas num tempo ligeiramente superior a zero.

O teorema pode ser enunciado da seguinte forma. Se *f(t)* e *df(t)/dt* são ambas transformáveis por Laplace e se $\lim_{t\to\infty} f(t)$ existe, então

$$f(0+) = \lim_{s\to\infty} sF(s)$$

Teorema da integração real
Se *f(t)* for de ordem exponencial, então a transformada de Laplace de $\int f(t)$ existe e é dado por

$$\mathcal{L}\left[\int \boldsymbol{f(t)dt}\right] = \frac{\boldsymbol{F(s)}}{\boldsymbol{s}} - \frac{\boldsymbol{f^{-1}(0)}}{\boldsymbol{s}} \quad \textbf{(4.3)}$$

Onde *F(s)* = $\mathcal{L}$ *[f(t)]* e $f^{-1}(0) = \int f(t)dt$ calculado em *t = 0*.

Neste teorema pode concluir-se que a integração no domínio do tempo converte-se em divisão no domínio *s*. Se o valor inicial do integral for zero, a transformada de Laplace do integral de *f(t)* é dado por *F(s)/s*.

Este teorema pode ser ligeiramente modificado para lidar com o integral definido de *f(t*). Se *f(t)* for de ordem exponencial, a transformada de Laplace do integral definido $\int_0^t f(t)dt$ é dado por:

$$\mathcal{L}\left[\int \boldsymbol{f(t)dt}\right] = \frac{\boldsymbol{F(s)}}{\boldsymbol{s}} \quad \textbf{(4.4)}$$

Onde *F(s)* = $\mathcal{L}$*[f(t)].*

Teorema da diferenciação complexa
Se *f(t)* é transformável por Laplace, então, excepto nos pólos de *F(s)*

$$\mathcal{L}[\boldsymbol{tf(t)}] = -\frac{\boldsymbol{d}}{\boldsymbol{ds}}\boldsymbol{F(s)}$$

Onde *F(s)* = $\mathcal{L}$ *[f(t)]*. Este teorema é conhecido teorema da diferenciação complexa. Além disso,

$$\mathcal{L}\left[\boldsymbol{t^2 f(t)}\right] = \frac{\boldsymbol{d^2}}{\boldsymbol{ds^2}}\boldsymbol{F(s)}$$

Em geral,

$$\mathcal{L}[\boldsymbol{t^n f(t)}] = (\boldsymbol{-1})^{\boldsymbol{n}}\frac{\boldsymbol{d^n}}{\boldsymbol{ds^n}}\boldsymbol{F(s)}$$

4.1.1.2 Função de Transferência

Em teoria de controlo, as chamadas **funções de transferência** são normalmente utilizadas para descrever a relação entre entradas e saídas (*input-output*) de processos/sistemas que podem ser descritas por equações diferenciais lineares invariantes no tempo.

A função de transferência de um sistema de equações diferenciais lineares invariantes no tempo é definida como a razão entre a transformada de Laplace da função de saída (função resposta) e a transformada de Laplace da função da entrada (função excitação) sob a hipótese de que todas as condições iniciais são nulas.

Considere-se o sistema linear invariante no tempo descrito pela seguinte equação diferencial:

$$a_0\frac{d^n y}{dt^n} + a_1\frac{d^{n-1}y}{dt^{n-1}} + \cdots + a_{n-1}\frac{dy}{dt} + a_n y = b_0\frac{d^m x}{dt^m} + b_1\frac{d^{m-1}x}{dt^{m-1}} + \cdots + b_{m-1}\frac{dx}{dt} + b_m x \quad (4.5)$$

Em que *y* é a saída do sistema (*output*) e *x* a entrada (*input*). A função de transferência do sistema obtém-se tomando-se as transformadas de Laplace de ambos os membros da equação (4.5), sob a hipótese de que todas as condições iniciais são nulas, ou seja

$$Função\ de\ transferência = G(s) = \left.\frac{\mathcal{L}[Saída]}{\mathcal{L}[Entrada]}\right|_{condi\,ções\ inici\,ais\ nulas} = \frac{Y(s)}{X(s)} = \frac{b_0 s^m + b_1 s^{m-1} + \cdots + b_{m-1}s + b_m}{a_0 s^n + a_1 s^{n-1} + \cdots + a_{n-1}s + a_n}$$

Com recurso ao conceito de função de transferência é possível representar a dinâmica de um sistema por equações algébricas no domínio *s*. Se a mais alta potência de s no denominador da função de transferência for igual a n, o sistema diz-se de *n*-ésima ordem.

Como exemplo, considere-se o exemplo típico usado em teoria de controlo, o sistema mecânico linear massa-mola-amortecedor (Figura 4-1). A força externa *u(t)* é a entrada do sistema e o deslocamento *x(t)* da massa, medido a partir da posição de equilíbrio na ausência da força externa, é a saída. Este é um sistema de entrada simples e saída simples, de segunda ordem.

Para se obter a função de transferência, executa-se os seguintes passos:

1. Escrever a equação diferencial para o sistema
2. Tomar a transformada de Laplace da equação diferencial, assumindo todas as condições iniciais nulas
3. Tomar a razão da saída X(s) para a entrada U(s).

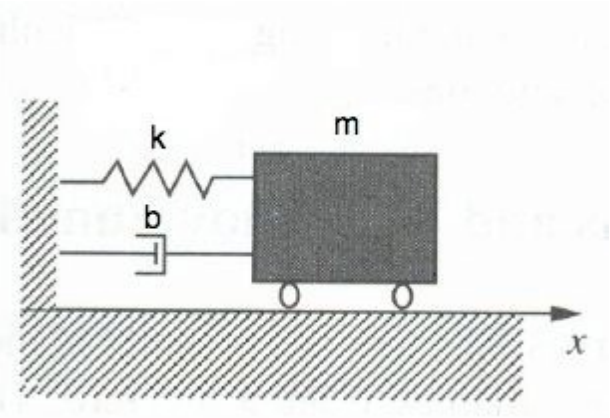


Figura 4-1 -Sistema massa-mola-amortecedror

A equação diferencial do sistema é

$$m\ddot{x} + b\dot{x} + kx = u$$

Tomando a transformada de Laplace de ambos os membros desta equação, admitindo todas as condições iniciais nulas, obtém-se

$$(ms^2 + bs + k)X(s) = U(s)$$

Onde *X(s) =* $\mathcal{L}$ *[x(t)]* e *U(s) =* $\mathcal{L}$*[u(t)].* A função de transferência será então dada por

$$\frac{X(s)}{U(s)} = \frac{1}{ms^2 + bs + k}$$

4.1.1.3 Função Resposta ao Impulso

Suponha-se que, para um sistema linear, invariante no tempo, a função de transferência é dada por

$$G(s) = \frac{Y(s)}{X(s)}$$

Onde $X(s)$ e $Y(s)$ são respectivamente as transformadas de Laplace da entrada e da saída do sistema. Então a saída $Y(s)$ pode ser formulada como o produto de $G(s)$ com $X(s)$

$$Y(s) = G(s)\,X(s) \tag{4.6}$$

Integral de Convulsão
A multiplicação no domínio complexo é equivalente à convulsão no domínio do tempo, e portanto a transformada inversa de Laplace da equação (4.6) é dada pelo integral de convulsão

$$y(t) = \int_0^t x(\tau)g(t-\tau)\tau$$

$$= \int_0^t g(\tau)x(t-\tau)\tau$$

Onde $g(t) = x(t)$ para $t < 0$.

Se $f_1(t)$ e $f_2(t)$ são funções bem comportadas continuas e de ordem exponencial, então

$$\boldsymbol{\mathcal{L}}\left[\int_0^t \boldsymbol{f_1}(t-\tau)\boldsymbol{f_2}(\tau)\right] = F_1(s)F_2(s) \tag{4.7}$$

onde

$$F_1(s) = \int_0^t \boldsymbol{f_1}(t)\boldsymbol{e^{-st}}\boldsymbol{dt} = \boldsymbol{\mathcal{L}}[\boldsymbol{f_1}(t)]$$

$$F_2(s) = \int_0^t \boldsymbol{f_2}(t)\boldsymbol{e^{-st}}\boldsymbol{dt} = \boldsymbol{\mathcal{L}}[\boldsymbol{f_2}(t)]$$

(Ogata, 1997)

Resposta ao impulso
Como a transformada de Laplace da função impulso unitário é a unidade, a transformada de Laplace da resposta ao impulso unitário de um sistema, quando as condições unitárias são nulas será simplesmente

$$Y(s) = G(s)$$

A respectiva transformada inversa de Laplace é a função resposta ao impulso, que é também a função de transferência do respectivo sistema

$$y(t) = \mathcal{L}^{-1}[G(s)] = g(t)$$

Esta função é também chamada a **função "peso"** do sistema. Assim, a função de transferência e a função resposta ao impulso unitário de um sistema linear invariante no tempo contém a mesma informação sobre a dinâmica do sistema. É possível assim obter a informação completa sobre as características dinâmicas do sistema.

Suponha-se que um sistema, cuja resposta ao impulso unitário é $g(t)$, tem uma entrada descrita por uma função $x(t)$. Suponha-se ainda que a entrada se inicia em $t = 0$ e dura até t_1. A entrada $x(t)$ pode ser aproximada por uma sequência de *N* funções pulso cuja duração é $\Delta t_1 = t_1/N$. Se Δt_1 é suficiente pequeno quando comparado com a menor constante de tempo do sistema, então 0 *k*-ésimo pulso pode ser considerado como um impulso cuja magnitude é a área $x(k\Delta t_1)\Delta t_1$. A resposta ao *k*-ésimo pulso no instante t será dada pelo produto da área do impulso e a função resposta ao impulso atrasada de $k\Delta t_1$

$$x(k\Delta t_1)\Delta t_1 g(t - k\Delta t_1)$$

Como o sistema é linear, aplica-se o princípio da sobreposição, pelo que a resposta $y(t)$, do sistema no instante *t*, à sequência das *N* funções pulso será dada pela soma de convulsão

$$y(t) = \sum_{0}^{N-1} x(k\Delta t_1) g(t - k\Delta t_1)\, \Delta t_1 \qquad \textbf{(4.8)}$$

onde $g(\tau) = 0$ para $\tau < 0$. Se tirarmos o limite, $\Delta t_1 \to 0$ (ou $N \to \infty$), então a resposta do sistema será dada pelo integral de convulsão

$$y(t) = \int_0^t x(t) g(t - \tau)\, d\tau \qquad \textbf{(4.9)}$$

4.1.1.4 Representação em espaço de estados

Uma representação em função de transferência (representação externa) permite encontrar uma expressão analítica que relaciona as duas variáveis do sistema, mas não descreve o que se passa em termos das outras variáveis externas.

A teoria de controlo moderno contrasta com a teoria convencional no sentido de que a primeira é aplicável a sistemas MIMO caracterizados por entradas múltiplas e saídas múltiplas (*multi-input/multi-output*), que podem ser lineares ou não-lineares, invariantes ou variantes no tempo. Esta nova abordagem (representação interna) é baseada no conceito de estado.

A ideia chave é que deverá ser possível encontrar um conjunto mínimo de variáveis que caracteriza o estado de um sistema dinâmico num determinado instante *t*. A evolução no tempo desses estados deverá ser suficiente para determinar completamente o comportamento do sistema (Botto, Controlo Óptimo, 2007).

Estado
O estado de um sistema é o menor conjunto de variáveis (chamadas **variáveis de estado**) tal que, o conhecimento destas variáveis no instante $t = t_0$, juntamente com o conhecimento das variáveis de entradas nos instantes $t \geq t_0$, determina completamente o comportamento do sistema em qualquer instante $t \geq t_0$.

Assim, o estado de um sistema dinâmico no instante t é univocamente determinado pelo estado no instante t_0 e pela entrada para $t \geq t_0$, e é independente do estado e da entrada antes de t_0 (Ogata, 1997).

O conceito de estado é aplicável não só a sistemas físicos, mas a qualquer outro tipo de sistema, nomeadamente, a sistemas biológicos, económicos, sociais etc.

Variáveis de estado
As variáveis de estado de um sistema dinâmico são as variáveis constituintes do conjunto mínimo de variáveis que determinam o estado do sistema dinâmico.

As variáveis de estado, por definição, não necessitam de ser mensuráveis ou observareis, e esta liberdade de escolha é uma vantagem do método do espaço de estados. No entanto, é de todo conveniente escolher, se possível, grandezas facilmente mensuráveis para variáveis de estado porque a estratégia de controlo óptimo requer a realimentação de todas as variáveis de estado.

Vector de estado
É o conjunto das *n* variáveis de estado necessárias para descrever completamente o comportamento de um dado sistema dispostas na forma vectorial. Um vector de estado é portanto um vector que determina univocamente o estado *x(t)* para qualquer instante $t \geq t_0$, uma vez conhecido o estado no instante $t = t_0$ e a entrada u(t) para $t \geq t_0$.

Espaço de estados $\boldsymbol{y(t)}$
Espaço *n*-dimensional constituído pelos eixos $x_1, x_2, \ldots, x_n$ onde podem ser representados os vectores de estados. Qualquer estado pode ser representado por um ponto no espaço de estados.

Equações do espaço de estados
No espaço de estados temos três tipos de variáveis:

1. Variáveis de entrada
2. Variáveis de saída
3. Variáveis de estado

A representação por espaço de estados para um determinado sistema não é única, mas o número de variáveis de estado é o mesmo para qualquer das diferentes representações do mesmo sistema.

A dinâmica do sistema deve envolver elementos que memorizem os valores de entrada para $t \geq t_1$, em sistemas de controlo contínuo, os integradores funcionam com dispositivos de memória, sendo que a saída de tais integradores podem ser consideradas como as variáveis que definem o estado interno da dinâmica do sistema. Nesta perspectiva, as saídas dos integradores comportam-se como variáveis de estado. O número total de variáveis que definem completamente a dinâmica do sistema é igual ao número de integradores que definem o sistema.

Considere-se um sistema MIMO com *n* integradores, contendo *r* entradas, $u_1(t), u_2(t) ,\ldots, u_r(t)$, e *m* saídas $y_1(t), y_2(t) ,\ldots, y_m(t)$. Definindo-se as *n* saídas dos integradores como variáveis de estado: $x_1(t), x_2(t) ,\ldots, x_n(t)$. Então o sistema pode ser descrito por

$$\begin{aligned} \dot{x}_1(t) &= f_1\left(x_1,\, x_2, \cdots, x_n, u_{1,} u_2 \,\cdots, u_r, t\right) \\ \dot{x}_2(t) &= f_2\left(x_1,\, x_2, \cdots, x_n, u_{1,} u_2 \,\cdots, u_r, t\right) \\ &\vdots \\ \dot{x}_n(t) &= f_n\left(x_1,\, x_2, \cdots, x_n, u_{1,} u_2 \,\cdots, u_r, t\right) \end{aligned} \qquad \textbf{(4.10)}$$

As saídas do sistema $y_1(t), y_2(t) ,\ldots, y_m(t)$ podem ser descritas na forma

$$\begin{aligned} y_1(t) &= g_1\left(x_1,\, x_2, \cdots, x_n, u_{1,} u_2 \,\cdots, u_r, t\right) \\ y_2(t) &= g_2\left(x_1,\, x_2, \cdots, x_n, u_{1,} u_2 \,\cdots, u_r, t\right) \\ &\vdots \\ y_n(t) &= g_m\left(x_1,\, x_2, \cdots, x_n, u_{1,} u_2 \,\cdots, u_r, t\right) \end{aligned} \qquad \textbf{(4.11)}$$

Definindo-se

$$\boldsymbol{x(t)} = \begin{bmatrix} \boldsymbol{x_1(t)} \\ \boldsymbol{x_2(t)} \\ \vdots \\ \boldsymbol{x_n(t)} \end{bmatrix}, \qquad f(x,,u,t) = \begin{bmatrix} f_1\left(x_1,\, x_2, \cdots, x_n, u_{1,} u_2 \,\cdots, u_r, t\right) \\ f_2\left(x_1,\, x_2, \cdots, x_n, u_{1,} u_2 \,\cdots, u_r, t\right) \\ \vdots \\ f_n\left(x_1,\, x_2, \cdots, x_n, u_{1,} u_2 \,\cdots, u_r, t\right) \end{bmatrix}$$

$$\boldsymbol{y(t)} = \begin{bmatrix} \boldsymbol{y_1(t)} \\ \boldsymbol{y_2(t)} \\ \vdots \\ \boldsymbol{y_m(t)} \end{bmatrix}, \qquad g(x,,u,t) = \begin{bmatrix} g_1\left(x_1,\, x_2, \cdots, x_n, u_{1,} u_2 \,\cdots, u_r, t\right) \\ g_2\left(x_1,\, x_2, \cdots, x_n, u_{1,} u_2 \,\cdots, u_r, t\right) \\ \vdots \\ g_m\left(x_1,\, x_2, \cdots, x_n, u_{1,} u_2 \,\cdots, u_r, t\right) \end{bmatrix}, \qquad \boldsymbol{u(t)} = \begin{bmatrix} \boldsymbol{u_1(t)} \\ \boldsymbol{u_2(t)} \\ \vdots \\ \boldsymbol{u_r(t)} \end{bmatrix}$$

Então as equações (4.10) e (4.11) podem ser escritas na forma compacta

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \qquad \textbf{(4.12)}$$

$$\mathbf{y}(t) = \mathbf{g}(\mathbf{x}, \mathbf{u}, t) \qquad \textbf{(4.13)}$$

Onde a equação (4.12) é a equação de estado e a equação (4.13) é a equação de saída (resposta). Se a função vector $\mathbf{f}$ e/ou $\mathbf{g}$ envolvem explicitamente o tempo, então o sistema é dito um “sistema variante com o tempo”.

Se as equações (4.12) e (4.13) forem linearizadas numa vizinhança do estado operativo, tem-se as equações de saída e de estado linearizadas

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) \quad \textbf{(4.14)}$$

$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) + \mathbf{D}(t)\mathbf{u}(t) \quad \textbf{(4.15)}$$

Onde $\mathrm{A}(t)$ é a matriz de estado, $\mathrm{B}(t)$ é a matriz de entradas, $\mathrm{C}(t)$, a matriz de saída e $\mathrm{D}(t)$ a matriz de transmissão directa. Um diagrama de blocos das equações (4.14) e (4.15) está representado na Figura 4-2.

Se as funções f e g não envolvem explicitamente o tempo, diz-se que o sistema é um sistema invariante no tempo. Neste caso, as equações (4.14) e (4.15) adquirem a forma

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \quad \textbf{(4.16)}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \quad \textbf{(4.17)}$$

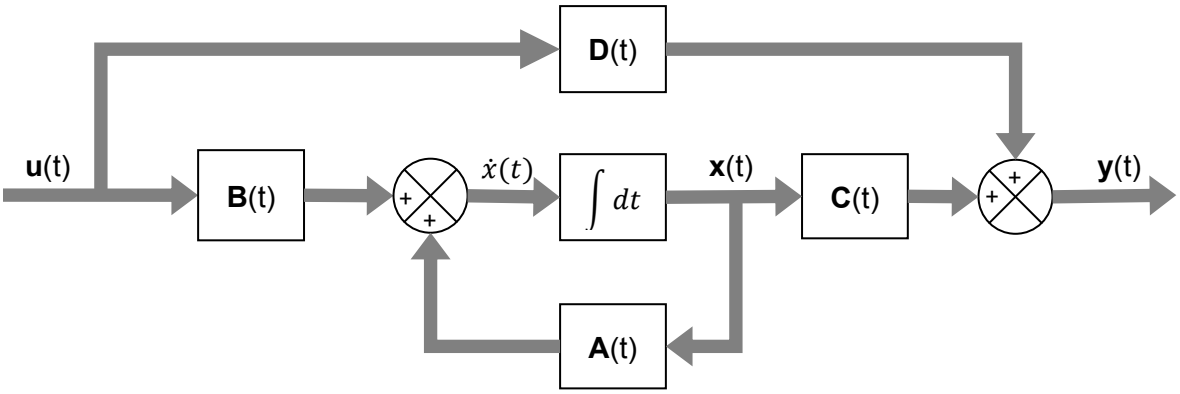


Figura 4-2 Diagrama de blocos de um sistema linear contínuo representado em espaço de estados

4.1.2 Sistemas discretos

Uma função temporal discreta $x_T(t)$ gerada a partir de uma função continua x(t) por amostragem periódica com um período de amostragem T_0 pode definir-se por

$$\begin{aligned} x_T(t) &= x(kT_0) && \textbf{para } \boldsymbol{t = kT_0} \\ x_T(t) &= 0 && \textbf{para } \boldsymbol{kT_0 < t < (k+1)T_0} \end{aligned} \qquad \boldsymbol{k = 0, 1, 2, \ldots} \quad \textbf{(4.18)}$$

Vários métodos de gerar funções discretas são apresentados nos exemplos que se seguem.

Considere-se a $x(t) = e^{-\alpha t}$. Se se retirarem valores desta função em intervalos de tempo constantes, $t = kT_0$, esta converte-se numa função discreta

$$x(kT_0) = e^{-\alpha k T_0} \qquad \boldsymbol{k = 0, 1, 2, \ldots}$$

Integração
O integral de uma função (não elementar)

$$x(t) = \frac{1}{T_1}\int_0^t w(t')dt'$$

Pode ser calculado numericamente por aproximação a uma função em escada:

$$x(kT_0) = \frac{1}{T_1}\sum_{v=0}^{k-1} T_0\, w(vT_0)$$

Neste caso, a área do rectângulo *k* é calculada multiplicando o valor (constante) do período de "amostragem" T_0, pelo valor da função no instante *k-1*, ou seja, $A_k = w(k\text{-}1)T_0$ (função escada e atraso).

Calculando este integral um passo à frente, obtém-se:

$$x\big((k+1)T_0\big) = \frac{1}{T_1}\sum_{v=0}^{k} T_0\, w(vT_0)$$

Subtraindo as duas equações resulta na relação recursiva

$$\boldsymbol{x\big((k+1)T_0\big) - x(kT_0) = \frac{T_0}{T_1}w(kT_0)}$$

Se substituirmos kT_0, se definirmos $a_1 = -1$ e $b_1 = T_0/T_1$, retardando um período, obtém-se:

$$\boldsymbol{x(k) + a_1 x(k-1) = b_1 w(k-1)}$$

Que é uma equação às diferenças de primeira ordem.

Se funções de tempo discreto, que dependem de outras funções de tempo discreto, podem ser escritas de forma recursiva, então obtém-se equações às diferenças. Uma equação às diferenças de *ordem m* é dada por

$$\begin{aligned}\boldsymbol{x(k) + a_1 x(k-1) + \cdots + a_m x(k-m)} \\ \boldsymbol{= b_0 w(k) + b_1 w(k-1) + \cdots + b_m w(k-m)}\end{aligned} \qquad \textbf{(4.19)}$$

O valor corrente da saída no instante t pode ser determinado recursivamente

$$\begin{aligned}\boldsymbol{x(k) = -a_1 x(k-1) - \cdots - a_m x(k-m) + b_0 w(k)} \\ \boldsymbol{+ b_1 w(k-1) + \cdots + b_m w(k-m)}\end{aligned} \qquad \textbf{(4.20)}$$

se forem conhecidos os valores das entradas (*inputs*) actuais e passados, e os valores das saídas (*outputs*) passadas.

4.1.2.1 Equações às diferenças de equações diferenciais

As equações às diferenças também podem ser obtidas pela discretização de equações diferenciais. Neste caso, um diferencial de primeira ordem é aproximado pela diferença de primeira ordem, um diferencial de segunda ordem pela diferença de segunda ordem, etc.

Para discretear equações diferenciais serão utilizadas as seguintes aproximações

$$\begin{aligned}
\frac{\boldsymbol{dx(t)}}{\boldsymbol{dt}} &\approx \frac{\boldsymbol{x(k)-x(k-1)}}{\boldsymbol{T_0}} \\
\frac{\boldsymbol{d^2x(t)}}{\boldsymbol{dt^2}} &\approx \frac{\boldsymbol{x(k)-2x(k-1)+x(k-2)}}{\boldsymbol{{T_0}^2}} \\
\frac{\boldsymbol{d^3x(t)}}{\boldsymbol{dt^3}} &\approx \frac{\boldsymbol{x(k)-3x(k-1)+3x(k-2)+x(k-3)}}{\boldsymbol{{T_0}^3}} \\
&\vdots \qquad\qquad\qquad\qquad \vdots
\end{aligned} \tag{4.21}$$

Discretização de um diferencial
Uma equação diferencial de primeira ordem é dada por

$$a_1 \frac{dx(t)}{dt} + x(t) = b_1 \mathrm{w(t)}$$

Aplicando (4.21) *obtém-se*

$$a_0 \mathrm{x(k)} + a_1 x(\mathrm{k}-1) = \mathrm{b}_0 \mathrm{w(k)}$$

Com

$$a_0 = \frac{a_1}{\mathrm{T}_0} + 1; \qquad \boldsymbol{a_1} = \frac{\boldsymbol{a_1}}{\mathbf{T_0}}; \qquad \boldsymbol{b_0} = \mathbf{b_0}$$

A equação (4.19) é a forma normalizada de uma equação às diferenças. A forma correspondente para uma equação diferencial resulta a partir da introdução das diferenças de ordem *n*

$$\begin{aligned}
&\alpha_n \Delta^n x(k) + \alpha_{n-1}\Delta^{n-1}x(k) + \cdots + \alpha_1 \Delta x(k) + x(k) \\
&\qquad = \beta_m \Delta^m w(k) + \cdots + \beta_1 w(k) + \beta_0 w(k)
\end{aligned}$$

4.1.2.2 Trem de impulsos

Um expediente matemático para tratamento de funções em tempo discreto obtém-se se o trem de pulsos $x_p(t)$ é aproximado por trem de impulsos $\delta - dirac$ (Isermann, 1989), onde a função $\delta - dirac$ é definido (ver acima Função impulso) por

$$\delta(t) = \begin{cases} 0, & t \neq 0 \\ \infty, & t = 0 \end{cases} \quad \textbf{(4.22)}$$

Deste modo, um sinal amostrado pode ser definido por um trem de impulsos

$$x^* = x(0)\delta(t) + x(T_0)\delta(t - T_0) + x(2T_0)\delta(t - 2T_0) + \cdots$$

ou

$$x^* = \sum_{k=0}^{\infty} x(kT_0)\delta(t - kT_0) \quad \textbf{(4.23)}$$

4.1.2.3 Transformada de Laplace de funções temporais discretas

A transformada de Laplace aplicada à função $\delta - dirac$ resulta (ver 4.1.1.1)

$$\mathcal{L}[\delta(t)] = \int_0^{\infty} \delta(t)\, e^{-st}\, dt = 1s \quad \textbf{(4.24)}$$

Se aplicado a impulsos deslocados

$$\mathcal{L}[\delta(t - kT_0)] = e^{-kT_0 s} \cdot 1s \quad \textbf{(4.25)}$$

Se aplicado a um trem de impulsos (4.23) (relembrar que para uma função transladada no tempo: $\mathcal{L}[f(t-\alpha)] = \int_0^{\infty} f(t-\alpha)\, e^{-st}\, dt = e^{-\alpha s} F(s)$ (Função transladada, acima))

$$\mathcal{L}[x^*(t)] = x^*(s) = \sum_{k=0}^{\infty} x(kT_0) e^{-kT_0 s} \cdot 1s \quad \textbf{(4.26)}$$

De notar que esta transformada de Laplace contém a dimensão *s.* A transformada de Laplace duma função discreta $x(kT_0)$ surge então periódica com frequência $\omega_0 = 2\pi/T_0$ (Isermann, 1989)

Teorema da amostragem de Shannon

Para cobrir um sinal continuo, de banda limitada, com uma frequência máxima ω_{max} com um sinal discreto, a frequência de amostragem terá que ser tal que (Isermann, 1989)

$$\omega_0 > 2\omega_{max} \implies T_0 < \pi/\omega_{max} \quad \textbf{(4.27)}$$

Deve ser mencionado que sinais contínuos de banda limitada, na realidade, não surgem apenas para sistemas de controlo. Contudo, em sistema de controlo a frequência de amostragem tem um papel fundamental, de modo a cobrir a dinâmica do sistema. Se, por exemplo, o período de amostragem for superior à constante de tempo do processo, o sinal

amostrado não evidencia a dinâmica do sistema, ou seja, não existe correlação entre sucessivos valores amostrados (Montgomery, 2001).

Elementos de retenção

Um amostrador é definido pela equação (4.18). Se um amostrador é seguindo por um retentor de ordem zero (ZOH - *zero order hold*), que retém o sinal amostrado $x(kT_0$ durante um período de amostragem, então resultará um sinal em escada (Figura 4-3).

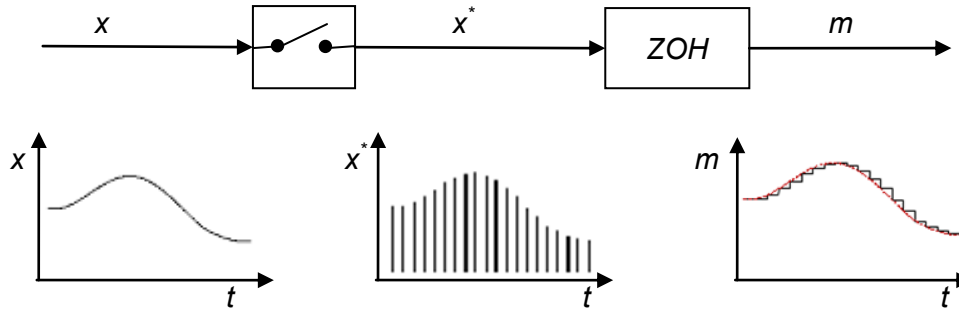


Figura 4-3 - Amostrador com zero-order hold

A função de transferência de um retentor de ordem zero é dada por (Isermann, 1989)

$$H(s) = \frac{m(s)}{x^*(s)} = \frac{1}{s}(1 - e^{-T_0 s}) \tag{4.28}$$

O ZOH pode ser modelado como integrador que tem apenas efeito durante um período. A sua resposta ao impulso corresponderá a um pulso de altura 1 e duração T_0.

4.1.2.4 Transformada Z

A transformada $\mathcal{Z}$ é a ferramenta ideal para lidar com sistemas discretos. A transformada $\mathcal{Z}$ da necessidade de eliminar o termo irracional $e^{-kT_0 s}$ na equação (4.26), que resulta sempre que se aplica a Transformada de Laplace a um sinal descrito por um trem de impulsos.

Se definirmos

$$z = e^{T_0 s} = e^{T_0(\delta + i\omega)} = e^{T_0 \delta}\left(cos(T_0\omega) + i\, sen(T_0\omega)\right) \tag{4.29}$$

$$z = e^{T_0 s} \quad \Leftrightarrow \quad s = \frac{1}{T_0}\ln(z)$$

Inserindo em (4.26), obtém-se a transformada $\mathcal{Z}$ de *x(t)*

$$X(z) = \mathcal{Z}[x(kT_0)] = \sum_{k=0}^{\infty} x(kT_0)z^{-k} \cdot 1s$$
$$= [x(0) + x(T_0)z^{-1} + x(2T_0)z^{-2} + \cdots] \cdot 1s$$

Resultando a definição geral de Transformada $\mathcal{Z}$ de um sinal:

$$Y(z) = \sum_{k=0}^{\infty} y(kT_0)z^{-k} \cdot 1s \quad \textbf{(4.30)}$$

De notar que, tal como a transformada de Laplace, a transformada $\mathcal{Z}$ também contém a dimensão s. A definição (4.30), apenas faz sentido se $Y(z)$ converge. Se $y(kT_0)$ é uma função limitada, $Y(z)$ converge para $|z| = |e^{T_0\delta}| > 1$.
Por vezes denomina-se ***z*** de operador de ***avanço*** e ***z-1*** o operador de ***atraso***, devido à associação: $z^{-1} \rightarrow \delta(t - T_0)$, $z^{-2} \rightarrow \delta(t - 2T_0)$, $\cdots$.

Geralmente, *Y(z)* é uma série infinita. Muitas funções podem contudo, ser representadas de forma fechada se usarmos as propriedades das series de potencias.

Função degrau
Suponha-se a função degrau:

$$Y_k = \begin{cases} a, & se\ k = 0,1,2,\cdots \\ 0, & se\ k = -1,-2,\cdots \end{cases} \Rightarrow$$

$$\Rightarrow \quad y^* = \delta(t) + a\delta(t - T_0) + a^2\delta(t - 2T_0) + \cdots$$

A transformada Z desta função resulta

$$Y(z) = 1 + \frac{a}{z} + \frac{a^2}{z^2} + \frac{a^3}{z^3} + \cdots = \sum_{k=0}^{\infty} a^k z^{-k}$$

Sendo esta ultima expressão uma série geométrica de razão a, pode ser descrita de acordo com a forma fechada

$$Y(z) = \frac{1}{1 - az^{-1}} = \frac{z}{z - a}$$

Neste caso, se $a = 1$, y* corresponde ao degrau unitário amostrado com periodo T_0, e portanto é possível escrever directamente da tabela de transformadas Z (Isermann, 1989):

$$\mathcal{Z}[1(t)] = \frac{z}{z - 1}$$

4.1.2.5 Teoremas da Transformada $\mathcal{Z}$

Seguidamente enumeram-se alguns teoremas da transformada $\mathcal{Z}$**.** As respectivas demonstrações podem ser consultadas em (Isermann, 1989) entre outros.

Linearidade

$$\mathcal{Z}[ax_1(k) + bx_2(k)] = \ a\mathcal{Z}[x_1(k)] + \ b\mathcal{Z}[bx_2(k)] \quad \textbf{(4.31)}$$

Mudança (*shifting*)
Mudança para a direita d períodos de amostragem T_0 (mudança no passado)

$$\mathcal{Z}[x(k-d)] = \ z^{-d}x(z) \qquad d \ \geq 0 \quad \textbf{(4.32)}$$

Mudança para a esquerda d períodos de amostragem T_0 (mudança no futuro)

$$\mathcal{Z}[x(k+d)] = \ z^{d}\left[x(z) - \sum_{q=0}^{d-1} x(q)z^{-q}\right] \qquad d \ \geq 0 \quad \textbf{(4.33)}$$

Na mudança para a esquerda os valores da função original não mudada $x(k)$ desaparecem para $k = 0,1, \dots, d-1$ porque a transformada $\mathcal{Z}$ apenas está definida para $q > 0$.

Exemplo (função degrau)

$$\mathcal{Z}[1(k+3)] = \ z^3\left[\frac{z}{z-1} - (1 + z^{-1} + z^{-2})\right] = \frac{z}{z-1}$$

Amortecimento

$$\mathcal{Z}[x(k)e^{-akT_0}] = \ x(ze^{aT_0}) \qquad d \ \geq 0 \quad \textbf{(4.34)}$$

Valor inicial

$$x(0) = \ \lim_{k\to 0} x(kT_0) = \ \lim_{z\to\infty} x(z)$$

Valor final

$$\lim_{k\to\infty} x(kT_0) = \ \lim_{z\to 1}\frac{z-1}{z}x(z) = \ \lim_{z\to 1}(z-1)x(z)$$

Apenas é válido se o valor final $x(\infty)$ existe.

4.1.2.6 A transformada $\mathcal{Z}$ inversa

Ao contrário da transformada de Laplace, a transformada $\mathcal{Z}$ inversa não é única. A transformada $\mathcal{Z}$ inversa aplicada a $F(z)$ resulta em

$$F(z) \xrightarrow{z^{-1}} \ f(kT_0) \ \neq f(t)$$

Pelo que apenas os valores múltiplos dos instantes de amostragem são coincidentes com os valores de $f(t)$ (Botto, Controlo de Sistemas, 2007)

Para derivar a transformada $\mathcal{Z}$ inversa, recorre-se ao facto de $X(z)$ e consequentemente X^* ser periódica com $\omega_0 = 2\pi/T_0$ e ainda simétrica em relação ao eixo real do plano *s*. Esta é a

razão pela qual $X^*(s)$ tem de ter a operação transformada inversa aplicada apenas ao domínio $\delta - i\pi/T_0 \leq s \leq \delta + i\pi/T_0$.

Existem vários métodos de aplicar a transformada $\mathcal{Z}$ inversa:

1. Método da expansão em fracções parciais
2. Método da divisão polinomial
3. Método da fórmula inversa

Os dois últimos são bastante trabalhosos. O método da expansão em fracções consiste na seguinte sequência de operações:

1. Dividir $F(z)$ por z: $F(z) \rightarrow F(z)/z$.
2. Expandir $F(z)/z$ em fracções parciais.
3. Multiplicar por z as fracções obtidas no ponto anterior.
4. Consultar a tabela de transformadas $\mathcal{Z}$ e aplicar a transformada inversa.

4.1.2.7 Função de transferência discreta

Para controlar os processos recorrer-se a medições periódicas às respectivas entrada e saída, o que implica a necessidade de introdução dos respectivos amostradores (*samplers*), como esquematizado na Figura 4-4.

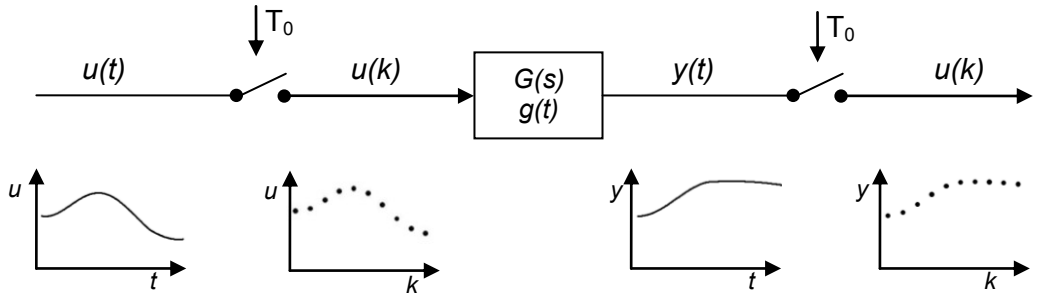


Figura 4-4 - Processo linear com amostradores de entrada e saída

Um amostrador à entrada de um sistema linear opera com função de transferência $G(s)$ ou resposta ao impulso $g(t)$. A entrada do sistema é então descrita pelo trem de impulsos

$$u^*(t) = \sum_{k=0}^{\infty} u(kT_0)\delta(t - kT_0) \tag{4.35}$$

Com a resposta ao impulso $g(t)$, a saída $y(t)$ será dada pelo somatório de convulsão (ver Resposta ao impulso)

$$y(t) = \sum_{k=0}^{\infty} u(kT_0)\delta(t - kT_0) \tag{4.36}$$

Se os amostradores de entrada e saída estão sincronizados, então, para $t = nT_0$ resulta

$$\begin{aligned} y(nT_0) &= \sum_{k=0}^{\infty} u(kT_0)\delta\big((n-k)T_0\big) \\ &= \sum_{v=0}^{\infty} u(n-v)\delta(vT_0) \end{aligned} \quad \textbf{(4.37)}$$

Tal como o integral de convulsão para sistemas contínuos, a saída do sistema no instante nT_0 é dada pelo somatório de convulsão de $u(kT_0)$ e $g\big((n-k)T_0\big)$.

Introduzindo a transformada de Laplace (4.26), à saída do sistema, $y(nT_0)$ descrita por (4.37), resulta

$$\begin{aligned} y^*(s) &= \sum_{n=0}^{\infty} y(kT_0)e^{-nT_0 s} \\ &= \sum_{n=0}^{\infty}\sum_{k=0}^{\infty} u(kT_0)\, g\big((n-k)T_0\big)e^{-nT_0 s} \end{aligned} \quad \textbf{(4.38)}$$

E substituindo $q = n - k \quad \Rightarrow \quad n = q + k$

$$\begin{aligned} y^*(s) &= \sum_{q=-k}^{\infty}\sum_{k=0}^{\infty} u(kT_0)\, g(qT_0)e^{-qT_0 s}\, e^{-kT_0 s} \\ &= \underbrace{\sum_{n=0}^{\infty} g(qT_0)e^{-qT_0 s}}_{G^*(s)}\ \underbrace{\sum_{k=0}^{\infty} u(kT_0)\, e^{-kT_0 s}}_{u^*(s)} \end{aligned} \quad \textbf{(4.39)}$$

Considerando que $g(qT_0) = 0$ para $q < 0$. Consequentemente, a transformada de Laplace $u^*(s)$ da entrada pode ser expresso como um factor separado e, tal como nos sistemas contínuos, a função de transferência será dada por

$$G^*(s) = \frac{y^*(s)}{u(s)} = \sum_{q=0}^{\infty} g(qT_0)e^{-qT_0 s} \quad \textbf{(4.40)}$$

Substituindo $z = e^{T_0 s}$ conduz-nos à função de transferência no domínio *z*

$$G(z) = \frac{Y(z)}{U(z)} = \sum_{q=0}^{\infty} g(qT_0)z^{-q} = z[g(q)] \quad \textbf{(4.41)}$$

Consequentemente, a função de transferência no domínio *z* é a relação entre a transformada z da saída amostrada e a transformada z da entrada amostrada que coincide com a transformada *z* da função resposta ao impulso.

Se o amostrador é seguido por um retentor (Figura 4-5), então tem-se que

$$HG(z) = \mathcal{L}[H(s)G(s)] \tag{4.42}$$

Para um retentor de ordem zero (ver Elementos de retenção)

$$\begin{aligned} HG(z) &= \mathcal{L}\left[\frac{1-e^{-T_0 s}}{s}G(s)\right] = \mathcal{L}\left[\frac{G(s)}{s}\right] - \mathcal{L}\left[\frac{G(s)}{s}e^{-T_0 s}\right] \\ &= (1-z^{-1})\,\mathcal{L}\left[\frac{G(s)}{s}\right] = \frac{z-1}{z}\mathcal{L}\left[\frac{G(s)}{s}\right] \end{aligned} \tag{4.43}$$

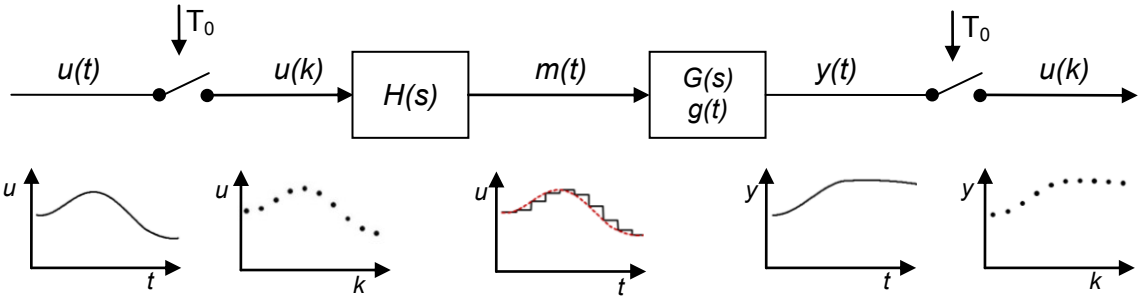


Figura 4-5 - Processo linear com retentor e amostradores de entrada e saída

Se a descrição de um sistema linear está na forma de equação às diferenças (equação (4.19)), então poderá ser escrita na forma

$$\begin{aligned} \boldsymbol{y(k) + a_1 y(k-1) + \cdots + a_m y(k-m)} \\ \boldsymbol{= b_0 w(k) + b_1 w(k-1) + \cdots + b_m w(k-m)} \end{aligned} \tag{4.44}$$

Aplicando directamente a propriedade de mudança (*shifting*) para o lado direito – atraso, tem-se

$$\boldsymbol{\left(1 + a_1 z^{-1} + \cdots + a_m z^{-m}\right) y(z) = \left(b_0 + b_1 z^{-1} + \cdots + b_m z^{-m}\right) u(z)}$$

Conduzindo-nos directamente para a função de transferência no domínio *z*

$$\boldsymbol{G(z) = \frac{y(z)}{u(z)} = \frac{b_0 + b_1 z^{-1} + \cdots + b_m z^{-m}}{1 + a_1 z^{-1} + \cdots + a_m z^{-m}} = \frac{B\left(z^{-1}\right)}{A(z^{-1})}} \tag{4.45}$$

4.1.2.8 Determinação da equação às diferenças a partir de G(z)

Por outro lado, de acordo com as tabelas da transformada z, a função de transferência genérica é dada na forma (Isermann, 1989):

$$\boldsymbol{G(z) = \frac{Y(z)}{U(z)} = \frac{b_0 z^n + b_1 z^{n-1} + \cdots + b_n}{z^n + a_1 z^{n-1} + \cdots + a_{n-1} z + a_n}}$$

Desenvolvendo, é possível obter a equação às diferenças correspondente

$$\boldsymbol{Y(z)\left(a_1 z^{n-1} + \cdots + a_{n-1}z + a_n\right) = U(z)\left(b_0 + b_1 z^{n-1} + \cdots + b_{n-1}z + b_n\right)}$$

Dividindo ambos os membros por z^n

$$\boldsymbol{Y(z) + a_1 z^{-1}Y(z) + \cdots + a_{n-1}z^{n-1}Y(z) + a_n z^{-n}Y(z) = b_0 U(z) + b_1 z^{-1}U(z) + \cdots + b_n z^{-n}U(z)}$$

Aplicando a propriedade de mudança (*shifiting*), $\mathcal{Z}[f(t - dT_0)] = z^{-d}F(z)$

$$\boldsymbol{y(t) + a_1 y(t - T_0) + \cdots + a_n y(t - nT_0) = b_0 u(t) + b_1 u(t - T_0) + \cdots + b_n\, u(t - nT_0)}$$

Onde $t = kT_0$, com $k = o, 1,2, \ldots$ correspondente aos instantes de amostragem. Se por outro lado considerar-se T_0 como um instante unitário($T_0 = 1$) então tem-se

$$\boldsymbol{y(t) + a_1 y(t-1) + a_2 y(t-2) + \cdots + a_n y(t-n) = b_0 u(t) + b_1 u(t-1) + \cdots + b_n\, u(t-n)}$$

Ou seja, a equação às diferenças genérica é dada por

$$\boldsymbol{y(t) = -a_1 y(t-1) - a_2 y(t-2) - \cdots - a_n y(t-n) + b_0 u(t) + b_1 u(t-1) + \cdots + b_n\, u(t-n)}$$

Introduzindo o operador de atraso $\mathcal{B}$ definido como: $\mathcal{B}^d\big(y(t)\big) = y(t-d)$, ou $\mathcal{B}^d Y_t = Y_{t-d}$ resulta

$$\boldsymbol{y(t) = -\left(a_1\mathcal{B} + a_2\mathcal{B}^2 + \cdots + a_n\mathcal{B}^n\right)y(t) + \left(b_0 + b_1\mathcal{B} + \cdots + b_n\,\mathcal{B}^n\right)u(t)}$$

$$\boldsymbol{Y_t = \frac{b_0 + b_1\mathcal{B} + \cdots + b_m\,\mathcal{B}^m}{1 + a_1\mathcal{B} + a_2\mathcal{B}^2 + \cdots + a_n\mathcal{B}^n}U_t = \frac{B(\mathcal{B})}{A(\mathcal{B})}U_t}$$

4.1.2.9 Propriedades da função de transferência z e equações às diferenças

Comportamento Proporcional

Para processos com comportamento proporcional, o ganho obtêm-se recorrendo ao teorema do valor final

$$K = \frac{y(k\to\infty)}{u(k\to\infty)} = \frac{\lim_{z\to1}(z-1)y(z)}{\lim_{z\to1}(z-1)u(z)} = \lim_{z\to1}\frac{y(z)}{u(z)}$$

$$= \lim_{z\to1} G(z) = \frac{b_0 + b_1 + \cdots + b_m}{1 + a_1 + \cdots + a_m} \qquad \textbf{(4.46)}$$

Comportamento Integral
No domínio s, o comportamento integral é descrito por $G(s) = 1/(T_1 s)$, directamente da tabela da transformada z verifica-se que $G(z) = \mathcal{L}\left[\frac{1}{T_1 s}\right] = \frac{1}{T_1(z-1)} = \frac{1}{T_1(1-z^{-1})}$. Ou seja, processos com comportamento integral têm um pólo em $z = 1$

$$\boldsymbol{G(z) = \frac{y(z)}{u(z)} = \frac{1}{1-z^{-1}} \frac{b_0 + b_1 z^{-1} + \cdots + b_m z^{-m}}{1 + a'_1 z^{-1} + \cdots + a'_m z^{-(m-1)}}} \quad \textbf{(4.47}$$

O gradiente "estado estacionário" após uma entrada em degrau de altura u_0 será descrito por

$$\begin{aligned} \lim_{k\to\infty} \Delta y(k) &= \lim_{k\to\infty} (y(k) - y(k-1)) \\ &= \lim_{z\to 1} y(z)\left(1 - z^{-1}\right) \\ &= \frac{b_0 + b_1 + \cdots + b_m}{1 + a'_1 + \cdots + a'_m} u_0 \end{aligned} \quad \textbf{(4.48)}$$

Se $b_0 \neq 0$ então o sistema tem um salto descontinuo em $k = 0$. Contudo, na maioria dos processos reais $b_0 = 0$ (Isermann, 1989).

Realizabilidade
Um sistema descrito por uma função de transferência discreta é realizável se o principio da causalidade é satisfeito, ou seja, se a variável de saída $y(k)$ não depende de valores futuros da entrada $u(k+j), j = 1,2,\cdots$. Assim, é necessário que a saída do sistema no instante $k, y(k)$, seja função apenas de valores passados da saída e de valores passados e/ou actuais da entrada, ou seja, a ordem do numerador de $G(s)$ tem que ser maior ou igual à ordem do denominador de $G(s)$.

4.1.2.10 Estabilidade

Considere-se o processo linear continuo descrito por

$$\begin{aligned} \boldsymbol{G(s)} &= \boldsymbol{\frac{y(s)}{u(s)} = \frac{B(s)}{A(s)} = \frac{b_0 + b_1 s + b_2 s^2 + \cdots + b_m s^m}{1 + a_1 s + a_2 s^2 + \cdots + a_n s^n}} \\ &= \frac{(s - s_{01})(s - s_{02}) \cdots (s - s_{0m})}{(s - s_1)(s - s_2) \cdots (s - s_{0n})} \frac{b_m}{a_n} \end{aligned} \quad (4.49)$$

Para o correspondente processo amostrado, de acordo com 3.4 de (Isermann, 1989), com ou sem ZOH, a função de transferência em z na forma racional será dada por

$$\boldsymbol{G(z) = \frac{y(z)}{u(z)} = \frac{B(z^{-1})}{A(z^{-1})} = \frac{b_0 + b_1 z^{-1} + \cdots + b_m z^{-m}}{1 + a_1 z^{-1} + \cdots + a_n z^{-n}}} \quad (4.50)$$

E após re-arranjo $(n > m)$

$$\boldsymbol{G(z)} = \frac{\boldsymbol{B(z^{-1})}}{\boldsymbol{A(z^{-1})}} = \frac{\boldsymbol{\left(b_0 z^m + b_1 z^{m-1} + \cdots + b_m\right) z^{n-m}}}{\boldsymbol{z^n + a_1 z^{n-1} + \cdots + a_n}}$$

$$= \frac{\boldsymbol{(z - z_{01})(z - z_{02}) \cdots (z - z_{0n})}}{\boldsymbol{(z - z_1)(z - z_2) \cdots (z - z_n)}} \quad (4.51)$$

As raízes $z_i,\ i = 1, 2, \ldots, n$ do polinómio do denominador $A(z) = 0$, são os pólos e as raízes $z_{0i},\ i = 1, 2, \ldots, n$ do polinómio do numerador $B(z) = 0$ são os zeros da função de transferência z.

Deste modo os pólos descrevem o comportamento modal do processo e dependem dos seus acoplamentos internos, pelo que são relevantes para a estabilidade. Os zeros indicam como é que a variável de entrada afecta as variáveis internas e como estas, por seu lado, influenciam a variável de saída (Isermann, 1989).

(Botto, Suplemento de Sebenta de Controlo de Sistemas, 2007) e (Isermann, 1989) mostram detalhadamente que um sistema discreto é estável se todos os pólos pertencerem ao interior do circulo unitário, isto é, se o modulo de todos os pólos de $G(z$ for inferior a 1 $\Rightarrow |z_j| < 1,\ \ j = 1, \cdots, n$

Se existir pelo menos um pólo sobre o circulo unitário com multiplicidade unitária, isto é, $|z_j| = 1$, e não existirem pólos instáveis, o sistema está no limite de estabilidade.

O sistema é instável se existir pelo menos um pólo fora do circulo unitário, ou seja, se existir um pólo com modulo superior a $1\left(|z_j| > 1\right)$ ou se existirem pólos sobre o circulo unitário com multiplicidade superior a 1.

4.1.2.11 Representação em espaço de estados

Para processos com comportamentos complicados e para processos multivariados, os modelos de espaço de estados são normalmente mais adequados. Ao contrário da representação por função de transferência, a representação por variáveis de estado, para além da relação entre a entrada e saída, descreve o comportamento e inter-relação das variáveis internas, funcionando como variáveis de memória do sistema. Existem vários métodos possíveis para determinar as variáveis de estado. Um dos métodos consiste na transformação do sistema continuo descrito por (4.16)(4.16) e (4.17) recorrendo à introdução de ZOH à entrada e amostradores à entrada e saída (Isermann, 1989). Seguidamente apresenta-se um método por substituição directa na equação às diferenças.

Se um modelo descrito por equações às diferenças existe para um determinado processo, então a representação de estados pode derivar-se pela introdução das variáveis de estado.

A partir da equação às diferenças (4.44) e substituindo *k* por *k*+*n* obtém-se

$$\boldsymbol{y(k+n) + a_1 y(k+n-1) + \cdots + a_n y(k)} = \boldsymbol{b_0 u(k+n) + b_1 u(k+n-1) + \cdots + b_n u(k)} \quad \textbf{(4.52)}$$

Para simplificar faz-se $m = n$. O caso para $m \neq n$ pode sempre ser tido em conta pondo ao parâmetros a zero.

Baseado nos seguintes pressupostos

1. No presente instante k, as saídas $y(k+n-1),\cdots,y(k+1)$ são utilizadas como variáveis de estado.
2. Com o incremento do instante k para $k+1$, $k+1$ para $k+2$ ..., ou seja $y(k)$ assume o valor de $y(k+1)$, $y(k+1)$ o valor de $y(k+2)$, etc. Conduzindo-nos a que $x_1(k+1)=x_2(k)$ e $x_2(k+1)=x_3(k)$

Pode Introduzir-se as seguintes variáveis de estado

$$\boldsymbol{y(k)} = \boldsymbol{x_1(k)} \tag{4.53}$$

$$\left.\begin{array}{lll} \boldsymbol{y(k+1)} & = \boldsymbol{x_2(k)} & = \boldsymbol{x_1(k+1)} \\ \boldsymbol{y(k+2)} & = \boldsymbol{x_3(k)} & = \boldsymbol{x_2(k+1)} \\ \vdots & \vdots & \vdots \\ \boldsymbol{y(k+n-1)} & = \boldsymbol{x_n(k)} & = \boldsymbol{x_{n-1}(k+1)} \\ \boldsymbol{y(k+1)} & & = \boldsymbol{x_n(k+1)} \end{array}\right\} \tag{4.54}$$

Substituindo (4.54) em (4.52) e fazendo $b_n=1$ e $b_0=b_1=\cdots=b_{n-1}=0$ obtem-se

$$\boldsymbol{y(k+n)=x_1(k+1)=a_1x_n(k)-a_2x_{n-1}(k)-\cdots-a_nx_1(k)+1u(k)} \tag{4.55}$$

A partir de (4.54) e (4.55) tira-se a equação vectorial às diferenças que consiste na representação de estado da equação às diferenças

$$\begin{bmatrix} \boldsymbol{x_1(k+1)} \\ \boldsymbol{x_2(k+1)} \\ \vdots \\ \boldsymbol{x_{n-1}(k+1)} \\ \boldsymbol{x_n(k+1)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{1} & \cdots & \boldsymbol{0} \\ & & \vdots & & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{1} \\ \boldsymbol{-a_n} & \boldsymbol{-a_{n-1}} & \boldsymbol{-a_{n-2}} & \cdots & \boldsymbol{-a_1} \end{bmatrix} \begin{bmatrix} \boldsymbol{x_1(k)} \\ \boldsymbol{x_2(k)} \\ \vdots \\ \boldsymbol{x_{n-1}(k)} \\ \boldsymbol{x_n(k)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \\ \boldsymbol{1} \end{bmatrix} \boldsymbol{u(k)} \tag{4.56}$$

E a respectiva equação de saída

$$\boldsymbol{y(k)} = [\boldsymbol{1}\ \boldsymbol{0}\ \cdots\ \boldsymbol{0}\ \boldsymbol{0}] \begin{bmatrix} \boldsymbol{x_1(k)} \\ \boldsymbol{x_2(k)} \\ \vdots \\ \boldsymbol{x_{n-1}(k)} \\ \boldsymbol{x_n(k)} \end{bmatrix} \tag{4.57}$$

Com a introdução de um vector de estado x, uma matriz sistema em tempo discreto, um vector de controlo $\boldsymbol{b}$ e um vector de saída $\boldsymbol{c}$

$$\boldsymbol{x(k+1)} = \boldsymbol{Ax(k)+b\,u(k)} \tag{4.58}$$

$$\boldsymbol{y(k)} = \boldsymbol{c^T x(k)} \tag{4.59}$$

Se as duas últimas equações são resolvidas de forma sequencial, a equação (4.59) poderá ser reescrita numa forma mais expedita

$$\boldsymbol{y(k+1) = c^T x(k+1)} \tag{4.60}$$

Estes resultados forma determinados com os pressupostos de que $b_n = 1$ e $b_0 = b_1 = \cdots = b_{n-1} = 0$, o que corresponde à função de transferência

$$\boldsymbol{y(z) = \frac{1}{z^n + a_1 z^{n-1} + \cdots + a_n} u(z) = x_1(z)} \tag{4.61}$$

Se, contudo, na equação (4.52), $b_n \neq 1$ e $b_0,\ b_1, \ldots, b_{n-1} \neq 0$, a equação (4.61) toma a forma

$$\begin{aligned} \boldsymbol{y(z)} &= \boldsymbol{\frac{b_0 z^n + b_1 z^{n-1} + \cdots + b_n}{z^n + a_1 z^{n-1} + \cdots + a_n} u(z)} \\ &= \boldsymbol{b_n x_1(z) + b_{n-1} z x_1(z) + \cdots + b_0 z^n x_1(z)} \end{aligned}$$

ou

$$\boldsymbol{y(k) = b_n x_1(k) + b_{n-1} x_1(k+1) + \cdots + b_0 x_1(k+n)} \tag{4.62}$$

Da equação (4.54) resulta

$$\boldsymbol{y(k) = b_n x_1(k) + b_{n-1} x_2(k) + \cdots + b_1 x_n(k) + b_0 x_n(k+1)} \tag{4.63}$$

Substituindo o termo $\boldsymbol{x_n(k+1)}$ dado pela equação (4.55) obtém-se finalmente

$$\begin{aligned} \boldsymbol{y(k)} &= \boldsymbol{b_n x_1(k) + b_{n-1} x_2(k) + \cdots + b_1 x_n(k)} \\ &\quad \boldsymbol{+ b_0 (a_1 x_n(k) - a_2 x_{n-1}(k) - \cdots - a_n x_1(k) + u(k))} \\ &= \boldsymbol{(b_n - b_0 a_n) x_1(k) + (b_{n-1} - b_0 a_{n-1}) x_2(k) + \cdots} \\ &\quad \boldsymbol{+ (b_1 - b_0 a_1) x_n(k) + b_0 u(k)} \end{aligned} \tag{4.64}$$

ou em notação vectorial

$$\boldsymbol{y(k) = [(b_n - b_0 a_n) \quad \cdots \quad (b_1 - b_0 a_1)] \begin{bmatrix} x_1(k) \\ \vdots \\ x_n(k) \end{bmatrix} + b_0 u(k)} \tag{4.65}$$

$$\boldsymbol{y(k) = c^T x(k) + d\, u(k)}$$

Para o caso em que $b_0 = 0$ esta equação toma a forma

$$\boldsymbol{y(k)} = [\boldsymbol{b_n} \quad \cdots \quad \boldsymbol{b_1}] \begin{bmatrix} \boldsymbol{x_1(k)} \\ \vdots \\ \boldsymbol{x_n(k)} \end{bmatrix}$$

Esta escolha de variáveis de estado é referida como “Forma canónica de controlabilidade”. É possível obter diversas representações em espaço de estado equivalentes, da forma

$$\begin{aligned} \boldsymbol{x(k+1)} &= \mathbf{A}\, \boldsymbol{x(k)} + \mathbf{B}\, \boldsymbol{u(k)} \\ \boldsymbol{y(k)} &= \boldsymbol{Cx(k)} + \boldsymbol{D}\, \boldsymbol{u(k)} \end{aligned}$$

Considerando a transformação linear e não singular, T, tal que $z(k) = T\, x(k)$, resulta

$$\begin{aligned} \boldsymbol{z(k+1)} &= \mathbf{A}_{\boldsymbol{T}}\, \boldsymbol{x(k)} + \mathbf{B}_{\boldsymbol{T}}\, \boldsymbol{u(k)} \\ \boldsymbol{y(k)} &= \boldsymbol{C_T x(k)} + \boldsymbol{D_T}\, \boldsymbol{u(k)} \end{aligned}$$

onde

$$\begin{aligned} \mathbf{A}_{\boldsymbol{T}} &= \boldsymbol{T\, AT^{-1}} \\ \mathbf{B}_{\boldsymbol{T}} &= \boldsymbol{TB} \\ \boldsymbol{C_T} &= \boldsymbol{CT^{-1}} \\ \boldsymbol{D_T} &= \boldsymbol{D} \end{aligned}$$

As formas canónicas possíveis são casos particulares de transformações lineares a que correspondem estruturas específicas para as matrizes A_T, B_T e C_T (Botto, Controlo Óptimo, 2007), (Isermann, 1989), salientando-se

1. Forma canónica de controlabilidade.
2. Forma canónica de observabilidade.
3. Forma canónica do controlador.
4. Forma canónica do observador.
5. Forma canónica de Jordan (ou modal).

4.1.2.12 Determinação do modelo do processo

Os modelos matemáticos do processo podem ser determinados com recurso a análises teóricas ou experimentais.

Na análise teórica os modelos são determinados pelas equações de balanço, equações de estado e leis fenomenológicas. Uma vez determinadas, em geral um sistema de equações diferenciais que conduzem a modelos teóricos de processos que estrutura e parâmetros determinados, podem então ser resolvidas explicitamente.

No caso da análise experimental dos processos o modelo matemático é determinado com recurso a medições. Os valores de entrada e saída são avaliados de modo a que se estabeleça um modelo matemático entre eles. O modelo pode ser não paramétrico, por exemplo uma função transiente da resposta em frequência na forma tabular, ou paramétrico, por exemplo uma equação diferencial ou equação às diferenças. Para os modelos não paramétricos recorre-se à avaliação das medições com recurso a análise de Fourier ou análise de correlação, ao passo que para os modelos paramétricos utilizam-se métodos de identificação tais resposta ao degrau ou resposta em frequência ou por estimação de parâmetros.

Para identificação de modelos paramétricos discretos, os métodos de estimação de parâmetros são especialmente adequados. Para processos invariantes no tempo, assumem-se modelos da forma

$$\boldsymbol{y(z)} = \underbrace{\frac{\boldsymbol{B(z^{-1})}}{\boldsymbol{A(z^{-1})}}\boldsymbol{z^{-d}u(z)}}_{\substack{\boldsymbol{modelo\ do} \\ \boldsymbol{processo}}} + \underbrace{\frac{\boldsymbol{D(z^{-1})}}{\boldsymbol{C(z^{-1})}}\boldsymbol{v(z)}}_{\substack{\boldsymbol{modelo\ dos} \\ \boldsymbol{disturbios}}} \tag{4.66}$$

Em que os parâmetros desconhecidos do processo e possivelmente também dos distúrbios são estimados baseado nas medidas de $u(k)$ e $y(k)$ (Isermann, 1989). Para estimar os parâmetros usam-se métodos tais como mínimos quadrados, variáveis instrumentais, máxima verosimilhança, na forma não recursiva ou recursiva.

4.2 Modelos de Series Temporais

No ponto 4.1 acima foi possível conclui-se que a dinâmica de processo pode ser modelada através de equações às diferenças. Pôde ainda concluir-se que os processos, quando amostrados com uma periodicidade adequada, reflectem a sua dinâmica através de dados autocorrelacionados, que por sua vez podem ser modelados com recurso a equações às diferenças. (Box & Jenkins, Time Series Analysis: Forecasting and Control, 1970) propôs a uma classificação para modelos estocásticos de equações às diferenças chamado ARIMA (*autoregressive integrated moving average*) composto por três componentes que correspondem às siglas “AR”, “I” e “MA”. Nesta classificação, a estrutura do modelo é identificada pela sigla ARIMA(*p,d,q*) onde *p* é a ordem da componente auto-regressiva, AR, do modelo, *q* é ordem da componente média móvel, MA, do modelo e *d* o grau de diferenciação necessário para se obter a estacionaridade do processo.

Um modelo ARIMA pode ser então descrito pela seguinte estrutura de equação às diferenças

$$\underbrace{\left[\boldsymbol{1 + a_1\mathcal{B} + \cdots + a_p\mathcal{B}^p}\right]}_{\substack{\boldsymbol{A(\mathcal{B})} \\ \boldsymbol{Parte\ AR}}} \underbrace{\boldsymbol{(1-\mathcal{B})^d}}_{\boldsymbol{Parte\ I}} \boldsymbol{y(t)} = \underbrace{\left[\boldsymbol{1 + c_1\mathcal{B} + \cdots + cb_q\mathcal{B}^q}\right]}_{\substack{\boldsymbol{C(\mathcal{B})} \\ \boldsymbol{Parte\ MA}}} \boldsymbol{\varepsilon(t)} \tag{4.67}$$

Em que $\boldsymbol{\varepsilon(t)}$ é considerado ruído branco $[\varepsilon(t) \sim N(0, \sigma\text{^}2\ )]$. De notar que esta é apenas uma variante, uma vez que a notação, a localização e consequentemente o sinal dos coeficientes podem mudar conforme os autores e/ou objectivos do desenvolvimento.

A designação “MA” resulta de se considerar neste modelo o ruído branco ponderado pelos coeficientes do polinómio $\boldsymbol{C(\mathcal{B})}$**.** Esta ponderação é entendida como uma média pesada de $\boldsymbol{\varepsilon(t)}$ ao longo do tempo, e portanto não nula, justificando assim a designação de “média móvel”.

Se o modelo não contiver uma ou duas das três componentes do modelo, o nome do termo e a correspondente ordem desaparecem do modelo. Por exemplo, um processo com uma estrutura ARIMA(0,1,1) pode designar-se por processo IMA(1,1). Estes modelos mostram-se bastante úteis para construir modelos de funções de transferência de sistemas que se pretendem controlar ou ajustar na presença de ruído correlacionado, possivelmente não correlacionado (Del Castillo, 2002).

4.2.1 Modelos Autoregressivos

Um modelo autoregressivo puro, ARIMA(*p,0,0*) ou AR(*p*), representado na forma (4.67), depois de retiradas as componentes “I” e “MA” apresenta a forma

$$\left(\mathbf{1} + \boldsymbol{a_1}\mathcal{B} + \cdots + \boldsymbol{a_p}\mathcal{B}^{\boldsymbol{p}}\right)\boldsymbol{y(t)} \;=\; \boldsymbol{\varepsilon(t)}$$

Rearranjando os termos e adaptando a notação (Del Castillo, 2002) e (Box, Jenkins, & Reinsel, 2008) em que $a \equiv -\phi$ e $\boldsymbol{y(t)} \equiv Y_t$

$$\boldsymbol{Y_t} = \boldsymbol{\xi} + \boldsymbol{\phi_1 Y_{t-1}} + \boldsymbol{\phi_2 Y_{t-2}} + \cdots + \boldsymbol{\phi_p Y_{t-p}} + \boldsymbol{\varepsilon_t}$$

Onde os $\boldsymbol{\phi_i}'s$ são os parâmetros e $\{\boldsymbol{\varepsilon_t}\}$ é uma sequência de ruído branco, ou seja, um processo estocástico. Este modelo pode também ser descrito na forma matricial

$$\boldsymbol{Y_t} = \underbrace{[\boldsymbol{\xi} \quad \boldsymbol{\phi_1} \quad \boldsymbol{\phi_2} \quad \cdots \quad \boldsymbol{\phi_p}]}_{\boldsymbol{\theta^T}} \cdot \underbrace{\begin{bmatrix} \mathbf{1} \\ \boldsymbol{Y_{t-1}} \\ \boldsymbol{Y_{t-2}} \\ \vdots \\ \boldsymbol{Y_{t-p}} \end{bmatrix}}_{\boldsymbol{\varphi}} + \boldsymbol{\varepsilon_t}$$

$$\textbf{vector dos parâmetros} \qquad\qquad \textit{regressor}$$

$$\boldsymbol{Y_t} = \boldsymbol{\theta^T} \cdot \boldsymbol{\varphi} + \boldsymbol{\varepsilon_t}$$

A diferença deste modelo para um modelo de regressão linear, em que $\boldsymbol{\xi}$ aparece no papel do parâmetro de intercepção, é que os regressores são variáveis desfasadas no tempo (*lagged*) do mesmo processo, daí o nome de modelo autoregressivo. Se os parâmetros $\boldsymbol{\phi_i}$ são de forma tal a que o processo tenha comportamento estacionário (sem tendência) então o valor de ξ coincide com a média do processo, ou seja, definindo$\widetilde{Y}_t = Y_t - \mu$, pode escrever-se

$$\boldsymbol{\widetilde{Y}_t} \equiv \boldsymbol{Y_t} - \boldsymbol{\mu} = \boldsymbol{\phi_1}\boldsymbol{\widetilde{Y}_{t-1}} + \boldsymbol{\phi_2}\boldsymbol{\widetilde{Y}_{t-2}} + \cdots + \boldsymbol{\phi_p}\boldsymbol{\widetilde{Y}_{t-p}} + \boldsymbol{\varepsilon_t} = \boldsymbol{A(\mathcal{B})}\boldsymbol{\widetilde{Y}_t} + \boldsymbol{\varepsilon_t} \qquad \textbf{(4.68)}$$

Definindo-se $A(\mathcal{B}) = 1 - \phi_1\mathcal{B} - \phi_2\mathcal{B}^2 - \cdots - \phi_p\mathcal{B}^p$, (4.69) pode ser escrita na forma

$$\boldsymbol{A(\mathcal{B})\tilde{Y}_t = \varepsilon_t} \quad \textbf{(4.69)}$$

De acordo com o ponto 4.1.2.10, fazendo a equivalência entre z^{-m} no domínio z e o operador $\mathcal{B}^m$ no domínio do tempo, um processo AR(*p*) é estacionário se e só se $\boldsymbol{A(\mathcal{B})}$ é estacionário, o que significa que todas as suas raízes se encontram estritamente fora do ciclo unitário no domínio plano complexo $\mathcal{B}$. Os critérios de estacionaridade estão directamente indexados ao posicionamento das raízes do polinómio característico no plano s, z ou $\mathcal{B}$ como está muito bem documentado em (Ogata, 1997), (Isermann, 1989) ou (Botto, Suplemento de Sebenta de Controlo de Sistemas, 2007).

Para se ter uma ideia intuitiva sobre esta matéria tome-se como exemplo o processo descrito por (4.66). Se se fizer $\boldsymbol{B(z^{-1}) = 0}$, $\boldsymbol{A(z^{-1}) = 0}$, $\boldsymbol{D(z^{-1}) = 1}$ e $\boldsymbol{C(z^{-1}) = 1 - \phi z^{-1}}$ obtém-se o processo AR(*p*)

$$\boldsymbol{y(z^{-1}) = \frac{1}{1 - \phi z^{-1}} v(z)} \quad \textbf{(4.70)}$$

Multiplicando ambos os membros por z

$$\boldsymbol{y(z^{-1}) = \frac{z}{z - \phi} v(z)}$$

De acordo com o ponto 4.1.2.10, o processo é estacionário se as raízes de $z - \phi = 0$ ou se encontram dentro do ciclo unitário, ou seja se $\phi \leq 0$.

Passando (4.70) para o domínio $\mathcal{B}$ o mesmo processo passa a ser descrito na forma

$$\boldsymbol{y_t = \frac{1}{1 - \phi \mathcal{B}} \varepsilon_t \implies (1 - \phi \mathcal{B}) y_t = \varepsilon_t \implies A(\mathcal{B}) y_t = \varepsilon_t} \quad \textbf{(4.71)}$$

Com $A(\mathcal{B}) = \boldsymbol{1 - \phi \mathcal{B}}$ e o único zero de $A(\mathcal{B})$ dado $\boldsymbol{1 - \phi \mathcal{B} = 0 \Rightarrow \mathcal{B} = 1/\phi}$**.** Ou seja, de acordo com o ponto 4.1.2.10, $\boldsymbol{\phi}$ tem que estar no ciclo unitário, pelo que $\boldsymbol{1/\phi}$ tem que ser maior que zero. Este critério pode ser facilmente visionado do seguinte modo: expandindo (4.71) tem-se sucessivamente

$$\boldsymbol{y_t = \phi \mathcal{B} y_t + \varepsilon_t = \phi y_{t-1} + \varepsilon_t} \quad \textbf{(4.72)}$$

$$\boldsymbol{y_{t-1} = \phi y_{t-2} + \varepsilon_{t-1}} \quad \textbf{(4.73)}$$

$$\boldsymbol{y_{t-2} = \phi y_{t-3} + \varepsilon_{t-2}}$$

Substituindo (4.73) em (4.72) sucessivamente

$$\boldsymbol{y_t = \phi(\phi y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t = \phi^2 y_{t-2} + \phi \varepsilon_{t-1} + \varepsilon_t}$$

$$\begin{aligned} \boldsymbol{y_t} &= \boldsymbol{\phi^2}(\boldsymbol{\phi y_{t-3}} + \boldsymbol{\varepsilon_{t-2}})\boldsymbol{y_{t-2}} + \boldsymbol{\phi\varepsilon_{t-1}} + \boldsymbol{\varepsilon_t} \\ &= \boldsymbol{\phi^3 y_{t-3}} + \boldsymbol{\phi^2\varepsilon_{t-2}} + \boldsymbol{\phi\varepsilon_{t-1}} + \boldsymbol{\varepsilon_t} \\ &\vdots \end{aligned}$$

$$\boldsymbol{y_t} = \boldsymbol{\phi^t y_0} + \boldsymbol{\phi^{t-1}\varepsilon_1} + \cdots + \boldsymbol{\phi\varepsilon_{t-1}} + \boldsymbol{\varepsilon_t} = \boldsymbol{\phi^t y_0} + \sum_{\boldsymbol{i=0}}^{\boldsymbol{t}} \boldsymbol{\phi^{t-i}\varepsilon_i}$$

Fazendo-se t tender para infinito e condições iniciais nulas$(y_0 = 0)$

$$\boldsymbol{y_t} = \sum_{\boldsymbol{i=0}}^{\boldsymbol{\infty}} \boldsymbol{\phi^{t-i}\varepsilon_i} = \sum_{\boldsymbol{i=0}}^{\boldsymbol{\infty}} \boldsymbol{\phi^i \mathcal{B}^i \varepsilon_t} \equiv \frac{\mathbf{1}}{\mathbf{1} - \boldsymbol{\phi\mathcal{B}}} \boldsymbol{\varepsilon_t}$$

Ou seja, está-se perante uma série geométrica que é convergente apenas se e só se $|\phi| < 1$.

4.2.1.1 Momentos de um processo AR(p)

O valor esperado de um processo AR(p) estacionário é dado por

$$\boldsymbol{E[Y_t]} = \boldsymbol{\mu} = \boldsymbol{E}\left[\boldsymbol{\xi} + \boldsymbol{\phi_1 Y_{t-1}} + \boldsymbol{\phi_2 Y_{t-2}} + \cdots + \boldsymbol{\phi_p Y_{t-p1}} + \boldsymbol{\varepsilon_t}\right]$$

Sendo ε_t ruído branco, o seu valor esperado é nulo. Se $E[Y_j] = \mu$, qualquer que seja j, e sendo ξ uma constante, então tem-se que

$$\boldsymbol{\mu} = \boldsymbol{\xi} + \boldsymbol{\phi_1 E[Y_{t-1}]} + \boldsymbol{\phi_2 E[Y_{t-2}]} + \cdots + \boldsymbol{\phi_p E}\left[\boldsymbol{Y_{t-p1}}\right] = \boldsymbol{\xi} + \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\phi_j \mu}$$

Resolvendo em ordem a μ, obtem-se

$$\boldsymbol{\mu} = \frac{\boldsymbol{\xi}}{\mathbf{1} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\phi_j}}$$

Autocovariância e Autocorrelação de um processo AR(p)

Para determinar a autocovariância tira-se o valor esperado do produto de cada elemento do processo$\tilde{Y}_t$ (4.68) por $\tilde{Y}_{t-k}$

$$\begin{aligned} \boldsymbol{\gamma_k} &= \boldsymbol{E[Y_t Y_{t-k}]} \\ &= \boldsymbol{E}\left[\left(\boldsymbol{\phi_1 \tilde{Y}_{t-1}} + \boldsymbol{\phi_2 \tilde{Y}_{t-2}} + \cdots + \boldsymbol{\phi_p \tilde{Y}_{t-p}} + \boldsymbol{\varepsilon_t}\right) \boldsymbol{\tilde{Y}_{t-k}}\right] \\ &= \boldsymbol{E}\left[\boldsymbol{\phi_1 \tilde{Y}_{t-1}\, \tilde{Y}_{t-k}}\right] + \boldsymbol{E}\left[\boldsymbol{\phi_2 \tilde{Y}_{t-2}\, \tilde{Y}_{t-k}}\right] + \cdots + \boldsymbol{E}\left[\boldsymbol{\phi_p \tilde{Y}_{t-p}\, \tilde{Y}_{t-k}}\right] + \boldsymbol{E}\left[\boldsymbol{\varepsilon_t\, \tilde{Y}_{t-k}}\right] \end{aligned}$$

Como o ultimo termo da direita é nulo, uma vez que se trata duma sequência de termos independentes e identicamente distribuídos (IID). Desta ultima equação e da definição de autocovariância (Autocovariância e Autocorrelação, pag.26) obtém-se a função de autocovariância

$$\boldsymbol{\gamma_k} = \boldsymbol{\phi_1\gamma_{k-1}} + \boldsymbol{\phi_2\gamma_{k-2}} + \cdots + \boldsymbol{\phi_p\gamma_{k-p}} \qquad k = 0, 1, 2, 3, \ldots \qquad \textbf{(4.74)}$$

Dividindo por γ_0 obtém-se a função autocorrelação

$$\boldsymbol{\rho_k} = \frac{\boldsymbol{\gamma_k}}{\boldsymbol{\gamma_0}} = \boldsymbol{\phi_1\rho_{k-1}} + \boldsymbol{\phi_2\rho_{k-2}} + \cdots + \boldsymbol{\phi_p\rho_{k-p}} \qquad k = 0, 1, 2, 3, \ldots \qquad \textbf{(4.75)}$$

Verifica-se que tanto a função autocorrelação como a função autocovariância obedecem à mesma estrutura de equação às diferenças. Para um processo estacionário, isto significa que a autocorrelação será eventualmente zero para valores de k relativamente elevados. Digamos que para processos estacionários AR(*p*), a função autocorrelação decai exponencialmente (*tails off*). De facto, o "decaimento" depende da solução da equação às diferenças para o processo AR(*p*), onde se pode frequentemente verificar um comportamento oscilatório, mas o decaimento de $|\rho_k|$ será contudo exponencial (Del Castillo, 2002).

As equações (4.75) para $k = 1, 2, 3, \ldots, p$ são conhecidas como equações de Yule-Walker. Estas equações são utilizadas para estimar os coeficientes $\phi's$ recorrendo a estimativas dos valores de autocorrelação $\boldsymbol{r_k} = \boldsymbol{\hat{\rho}_k}$. Reescrevendo todas as equações (4.75), substituindo os valores da autocorrelação pelas estimativas e tendo em conta que $\boldsymbol{\rho_k} = \boldsymbol{\rho_{-k}}$, tem-se

$$\begin{aligned}
r_1 &= \boldsymbol{\phi_1 r_0} + \boldsymbol{\phi_2 r_1} + \cdots + \boldsymbol{\phi_{p-1} r_{p-2}} + \boldsymbol{\phi_p r_{p-1}} \\
\boldsymbol{r_2} &= \boldsymbol{\phi_1 r_1} + \boldsymbol{\phi_2 r_0} + \cdots + \boldsymbol{\phi_{p-1} r_{p-3}} + \boldsymbol{\phi_p r_{p-2}} \\
&\vdots \qquad\qquad\qquad\qquad \vdots \\
r_{p-1} &= \boldsymbol{\phi_1 r_{p-2}} + \boldsymbol{\phi_2 r_{p-3}} + \cdots + \boldsymbol{\phi_{p-1} r_0} + \boldsymbol{\phi_p r_1} \\
\boldsymbol{r_p} &= \boldsymbol{\phi_1 r_{p-1}} + \boldsymbol{\phi_2 r_{p-2}} + \cdots + \boldsymbol{\phi_{p-1} r_1} + \boldsymbol{\phi_p r_0}
\end{aligned}$$

Que pode ser escrito na forma compacta

$$\begin{bmatrix} \boldsymbol{r_1} \\ \boldsymbol{r_2} \\ \vdots \\ \boldsymbol{r_{p-1}} \\ \boldsymbol{r_p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{r_0} & \boldsymbol{r_1} & \cdots & \boldsymbol{r_{p-2}} & \boldsymbol{r_{p-1}} \\ \boldsymbol{r_1} & \boldsymbol{r_0} & \cdots & \boldsymbol{r_{p-3}} & \boldsymbol{r_{p-2}} \\ \vdots & \vdots & & \vdots & \vdots \\ \boldsymbol{r_{p-2}} & \boldsymbol{r_{p-3}} & \cdots & \boldsymbol{r_0} & \boldsymbol{r_1} \\ \boldsymbol{r_{p-1}} & \boldsymbol{r_{p-2}} & \cdots & \boldsymbol{r_1} & \boldsymbol{r_0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\phi_1} \\ \boldsymbol{\phi_2} \\ \vdots \\ \boldsymbol{\phi_{p-1}} \\ \boldsymbol{\phi_p} \end{bmatrix}$$

Relembrando que $r_0 = 1$ resulta

$$\underbrace{\begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_{p-1} \\ r_p \end{bmatrix}}_{r} = \underbrace{\begin{bmatrix} 1 & r_1 & \cdots & r_{p-2} & r_{p-1} \\ r_1 & 1 & \cdots & r_{p-3} & r_{p-2} \\ \vdots & \vdots & & \vdots & \vdots \\ r_{p-2} & r_{p-3} & \cdots & 1 & r_1 \\ r_{p-1} & r_{p-2} & \cdots & r_1 & 1 \end{bmatrix}}_{R} \underbrace{\begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{p-1} \\ \phi_p \end{bmatrix}}_{\Phi}$$

ou de forma sucinta

$$R\Phi = r \tag{4.76}$$

Note-se que se trata de sistema bem comportado com a matriz R dos coeficientes quadrada e vector das incógnitas Φ de rank completo. Mais ainda, R é uma matriz simétrica e de rank completo, de modo que a invertibilidade está garantida. Tem-se então que uma estimativa dos coeficientes $\phi's$ pode ser determinada através do cálculo

$$\hat{\Phi} = R^{-1} r \tag{4.77}$$

Variância de um processo AR(*p*)
A variância de um processo AR(p) obtém-se através da equação **(4.74)** fazendo $k = 0$

$$\gamma_0 = \phi_1\gamma_{-1} + \phi_2\gamma_{-2} + \cdots + \phi_p\gamma_{-p} + E\left[\varepsilon_t \tilde{Y}_t\right]$$

Onde, neste caso, o ultimo termo contém o valor esperado do produto de duas variáveis aleatórias que tem uma variável de erro comum (ε_t), pelo que será diferente de zero:

$$\gamma_0 = {\sigma_Y}^2 = \phi_1\gamma_1 + \phi_2\gamma_2 + \cdots + \phi_p\gamma_p + {\sigma_\varepsilon}^2 = \sum_{j=1}^{p} \phi_j\gamma_j + {\sigma_\varepsilon}^2$$

Dividindo ambos os membros por $\gamma_0 = {\sigma_Y}^2$ obtém-se após manipulação

$${\sigma_Y}^2 = \gamma_0 = \frac{{\sigma_\varepsilon}^2}{1 - \sum_{j=1}^{p} \phi_j\gamma_j} \tag{4.78}$$

4.2.1.2 Função Autocorrelação Parcial

Quando se tenta identificar um processo a partir de uma série temporal, numa primeira fase, será necessário estimar a ordem p do processo autoregressivo. A função de autocorrelação parcial é um dispositivo que explora o facto de, apesar de um processo AR(*p*) ter uma função autocorrelação infinita, poder ser descrita em termos de *p* funções de autocorrelação.

Definindo-se ϕ_{jk} como o *j*-ésimo coeficiente numa representação autoregressiva de ordem k, tal que ϕ_{kk} é o ultimo coeficiente. A partir da equação (4.75), para uma representação de ordem k, os coeficientes ϕ_{jk} satisfazem o conjunto de equações

$$\boldsymbol{\rho_j = \phi_{k1}\rho_{j-1} + \phi_{k2}\rho_{j-2} + \cdots + \phi_{k(k-1)}\rho_{j-k+1} + \phi_{kk}\rho_{j-k}} \quad k = 1, 2, \ldots, k \qquad \textbf{(4.79)}$$

Conduzindo-nos às equações de Yule-Walker

$$\underbrace{\begin{bmatrix} \mathbf{1} & \boldsymbol{\rho_1} & \boldsymbol{\rho_2} & \cdots & \boldsymbol{\rho_{k-1}} \\ \boldsymbol{\rho_1} & \mathbf{1} & \boldsymbol{\rho_1} & \cdots & \boldsymbol{\rho_{k-2}} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{\rho_{k-1}} & \boldsymbol{\rho_{k-2}} & \boldsymbol{\rho_{k-3}} & \cdots & \mathbf{1} \end{bmatrix}}_{\mathrm{P}} \underbrace{\begin{bmatrix} \boldsymbol{\phi_{k1}} \\ \boldsymbol{\phi_{k2}} \\ \vdots \\ \boldsymbol{\phi_{kk}} \end{bmatrix}}_{\boldsymbol{\phi}} = \underbrace{\begin{bmatrix} \boldsymbol{\rho_1} \\ \boldsymbol{\rho_2} \\ \vdots \\ \boldsymbol{\rho_k} \end{bmatrix}}_{\boldsymbol{\rho}} \qquad \textbf{(4.80)}$$

Resolvendo estas equações para $\boldsymbol{k = 1, 2, 3}$, ... sucessivamente para ϕ_{kk}, obtém-se

$$\boldsymbol{\phi_{11} = \rho_1}$$

$$\begin{bmatrix} \mathbf{1} & \boldsymbol{\rho_1} \\ \boldsymbol{\rho_1} & \mathbf{1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\phi_{21}} \\ \boldsymbol{\phi_{22}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\rho_1} \\ \boldsymbol{\rho_2} \end{bmatrix} \Rightarrow \boldsymbol{\phi_{22}} = \frac{\begin{vmatrix} \mathbf{1} & \boldsymbol{\rho_1} \\ \boldsymbol{\rho_2} & \boldsymbol{\rho_2} \end{vmatrix}}{\begin{vmatrix} \mathbf{1} & \boldsymbol{\rho_1} \\ \boldsymbol{\rho_1} & \mathbf{1} \end{vmatrix}} = \frac{\boldsymbol{\rho_2 - \rho_1^2}}{\boldsymbol{1 - \rho_1^2}}$$

$$\boldsymbol{\phi_{33}} = \frac{\begin{vmatrix} \mathbf{1} & \boldsymbol{\rho_1} & \boldsymbol{\rho_1} \\ \boldsymbol{\rho_1} & \mathbf{1} & \boldsymbol{\rho_2} \\ \boldsymbol{\rho_2} & \boldsymbol{\rho_1} & \boldsymbol{\rho_3} \end{vmatrix}}{\begin{vmatrix} \mathbf{1} & \boldsymbol{\rho_1} & \boldsymbol{\rho_2} \\ \boldsymbol{\rho_1} & \mathbf{1} & \boldsymbol{\rho_1} \\ \boldsymbol{\rho_2} & \boldsymbol{\rho_1} & \mathbf{1} \end{vmatrix}}$$

Ou seja, em geral, para ϕ_{kk} o determinante do numerador tem os mesmos elementos que o determinante do denominador, $det(\mathrm{P})$, mas em que a ultima coluna de P foi substituída por ρ, ver a equação **(4.80)**:

$$\boldsymbol{\phi_{kk}} = \frac{\begin{vmatrix} \mathbf{1} & \boldsymbol{\rho_1} & \cdots & \boldsymbol{\rho_{k-2}} & \boldsymbol{\rho_1} \\ \boldsymbol{\rho_1} & \mathbf{1} & \cdots & \boldsymbol{\rho_{k-3}} & \boldsymbol{\rho_2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{\rho_{k-1}} & \boldsymbol{\rho_{k-2}} & \cdots & \boldsymbol{\rho_1} & \boldsymbol{\rho_k} \end{vmatrix}}{\begin{vmatrix} \mathbf{1} & \boldsymbol{\rho_1} & \boldsymbol{\rho_2} & \cdots & \boldsymbol{\rho_{k-1}} \\ \boldsymbol{\rho_1} & \mathbf{1} & \boldsymbol{\rho_1} & \cdots & \boldsymbol{\rho_{k-2}} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{\rho_{k-1}} & \boldsymbol{\rho_{k-2}} & \boldsymbol{\rho_{k-3}} & \cdots & \mathbf{1} \end{vmatrix}} \qquad \textbf{(4.81)}$$

A quantidade $\boldsymbol{\phi_{kk}}$ vistas como funções da *lag k*, denomina-se como "Função Autocorrelação Parcial". Para um processo autoregressivo de ordem p, a função autocorrelação parcial $\boldsymbol{\phi_{kk}}$ será diferente de zero para *k* menor ou igual a *p* e zero para *k* maior que *p*. Por outras palavras, a função autocorrelação parcial de um processo autoregressivo de ordem *p* corta (é igual a zero) após a *lag p.*

Os $\boldsymbol{\phi_{kk}}$ são interpretados como a medida de correlação entre Y_t e Y_{t-k+1} após descontar o efeito das variáveis intermédiasY_{t-2}, Y_{t-2}, ... , Y_{t-k+1} (Del Castillo, 2002).

A função autocorrelação parcial dada através de **(4.80)** está definida para qualquer processo estacionário como uma função da autocorrelação ρ_k do processo, mas a característica

distintiva de que $\boldsymbol{\phi_{kk}} = \mathbf{0}$ para $k > p$ num processo AR(*p*) serve para caracterizar a ordem *p* do processo. A partir da teoria dos mínimos quadrados, pode estabelecer-se que os valores $\boldsymbol{\phi_{k1}}, \boldsymbol{\phi_{k1}}, \dots, \boldsymbol{\phi_{kk}}$, solução de (4.80), são os coeficientes de regressão numa regressão linear de Y_t em Y_{t-2}, Y_{t-2}, … , Y_{t-k}, isto é, são os valores dos coeficientes $b_1, \dots, b_k$ que minimizam $E\left[\left(z_t - b_0 - \sum_{i=1}^{k} b_i z_{t-i}\right)^2\right]$. Assim, assumindo por conveniência que o processo $\{Y_t\}$ tem média zero, o melhor predictor linear, no sentido dos mínimos desvios quadrados, de Y_t baseado em Y_{t-2}, Y_{t-2}, … , Y_{t-k} será (Box, Jenkins, & Reinsel, 2008)

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{t}} = \boldsymbol{\phi_{k-1,1}} \boldsymbol{Y_{t-k+1}} + \boldsymbol{\phi_{k-1,2}} \boldsymbol{Y_{t-k+2}} + \cdots + \boldsymbol{\phi_{k-1,k-1}} \boldsymbol{Y_{t-1}}$$

seja o processo AR ou não.

Estimativa da Função Autocorrelação Parcial

A estimativa da autocorrelação parcial (PACF – *Parcial Autocorrelation Function*) pode ser obtida pela substituição dos valores teóricos da autocorrelação ρ_j pelas respectivas estimativas r_j. A partir da equação **(4.76)** pode-se construir um método recursivo para cálculo das PACF. O primeiro passo consiste no cálculo das ACF (função autocorrelação, ver Autocovariância e Autocorrelação) até a uma razoável ordem de corte (*cutoff*), por exemplo $p \simeq N/4$. Seguidamente, seja $r^{(i)}$ o vector da equação **(4.76)** para o caso de $p = i$. De forma similar, seja $R^{(i)}$ a matriz dos coeficientes para o mesmo caso. Então, para determinar $PACF(i)$ como função de i, implementar o seguinte algoritmo:

1. Ciclo para $i = 1$ até $i = p$, incrementando uma unidade:
2. Calcular $R^{(i)}$ e $r^{(i)}$
3. Inverter $R^{(i)}$
4. $\hat{\Phi}^{(\mathrm{i})} = \left(R^{(i)}\right)^{-1} r^{(i)} = \begin{pmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \vdots \\ \hat{\phi}_i \end{pmatrix}$
5. Ignorar todos os $\hat{\phi}_j$ para $1 \leq j \leq i-1$
6. Reter $\hat{\phi}_i$
7. $\hat{\phi}_{ii} \equiv PACF(i) = \hat{\phi}_i$
8. Fim de ciclo em i

Desvio Padrão da estimativa da Autocorrelação Parcial

Na hipótese de que um processo é autoregressivo de ordem p, as autocorrelações parciais (PACF) estimadas de ordem $p+1$, e acima, são aproximadamente independentes e normalmente distribuídas com média zero, então

$$\boldsymbol{Var}\left[\widehat{\boldsymbol{\phi}}_{\boldsymbol{kk}}\right] \simeq \frac{\mathbf{1}}{\boldsymbol{n}} \qquad \boldsymbol{k} \geq \boldsymbol{p} + \mathbf{1}$$

Onde n é o numero de observações da série que serviu de base para o ajustamento (Box, Jenkins, & Reinsel, 2008). Assim, e de acordo com o ponto (Autocovariância e Autocorrelação), o desvio padrão (*DP*) da autocorrelação parcial estimada será dado por

$$\boldsymbol{DP}\left[\widehat{\boldsymbol{\phi}}_{\boldsymbol{kk}}\right] = \widehat{\boldsymbol{\sigma}}\left[\widehat{\boldsymbol{\phi}}_{\boldsymbol{kk}}\right] \simeq \frac{\mathbf{1}}{\sqrt{\boldsymbol{n}}} \qquad \boldsymbol{k} \geq \boldsymbol{p}+\mathbf{1} \tag{4.82}$$

4.2.2 Modelos Média Móvel

Se um processo pode ser representado por uma média movel ponderada de choques aleatórios IID $\{\varepsilon_t\}$ (ruido branco), então está-se perante um "Processo Média Movel" designado pela abreviatura MA.

Um modelo Média Movel puro, ARIMA(*0,0,q*) ou MA(*q*), representado na forma (4.67), depois de retiradas as componentes "AR" e "I" apresenta a forma

$$\boldsymbol{y}(\boldsymbol{t}) = \left(\mathbf{1} + \boldsymbol{c_1}\boldsymbol{\mathcal{B}} + \cdots + \boldsymbol{c_q}\boldsymbol{\mathcal{B}}^{\boldsymbol{q}}\right)\boldsymbol{\varepsilon}(\boldsymbol{t})$$

Rearranjando os termos e adaptando a notação (Del Castillo, 2002) em que $b \equiv -\theta$ e $\boldsymbol{y}(\boldsymbol{t}) \equiv Y_t$, tem-se que um processo MA(q) pode ser descrito por

$$\boldsymbol{Y_t} = \boldsymbol{\mu} - \boldsymbol{\theta_1}\boldsymbol{\varepsilon_{t-1}} - \boldsymbol{\theta_2}\boldsymbol{\varepsilon_{t-2}} - \cdots - \boldsymbol{\theta_q}\boldsymbol{\varepsilon_{t-q}} + \boldsymbol{\varepsilon_t} \tag{4.83}$$

Utilizando novamente a definição $\tilde{Y}_t = Y_t - \mu$, um processo MA pode ser descrito por

$$\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}} = \left(\mathbf{1} - \boldsymbol{\theta_1}\boldsymbol{\mathcal{B}} - \boldsymbol{\theta_2}\boldsymbol{\mathcal{B}}^{\mathbf{2}} - \cdots - \boldsymbol{\theta_q}\boldsymbol{\mathcal{B}}^{\boldsymbol{q}}\right)\boldsymbol{\varepsilon_t} \equiv \boldsymbol{C}(\boldsymbol{\mathcal{B}})\boldsymbol{\varepsilon_t} \tag{4.84}$$

Comparando um processo MA(*q*) puro com um AR(*p*) puro, verifica-se que o processo MA(q) tem um polinómio $A(\mathcal{B})$ igual a 1 e um processo AR(*p*) tem um polinómio $\boldsymbol{C}(\mathcal{B})$ igual a 1. Deste modo, um processo MA será sempre estacionário uma vez que o polinómio $A(\mathcal{B})$ não depende de $\mathcal{B}$ pelo que não terá qualquer raiz nem dentro nem fora do ciclo unitário.

Invertibilidade

Invertibilidade é uma condição matematicamente similar à estacionaridade relativamente ao polinómio $\boldsymbol{C}(\mathcal{B})$ de um processo MA. A condição de invertibilidade é diferente da condição de estacionaridade e tanto se aplica a modelos estacionários como a modelos não estacionários. Um processo ARIMA é invertível se este pode ser descrito como um processo AR da forma (Del Castillo, 2002)

$$\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}} \equiv \boldsymbol{\pi_1}\widetilde{\boldsymbol{Y}}_{\boldsymbol{t-1}} + \boldsymbol{\pi_2}\widetilde{\boldsymbol{Y}}_{\boldsymbol{t-2}} + \cdots + \boldsymbol{\pi_p}\widetilde{\boldsymbol{Y}}_{\boldsymbol{t-p}} + \boldsymbol{\varepsilon_t} \tag{4.85}$$

Onde a soma finita ou infinita mas tem que convergir para um valor finito. A invertibilidade significa que a influência das observações prévias $\widetilde{\boldsymbol{Y}}_{\boldsymbol{t-j}}$ em $\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}$ decrescem com a idade. Para ilustrar a ideia, considere-se o seguinte modelo MA(1)

$$\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}} = (\mathbf{1} - \boldsymbol{\theta}\boldsymbol{\mathcal{B}})\boldsymbol{a}_{\boldsymbol{t}} \tag{4.86}$$

Invertendo, ou seja, resolvendo para a_t em termos das observações passadas e presente de $\widetilde{\boldsymbol{Y}}$

$$\boldsymbol{a}_{\boldsymbol{t}} = (\mathbf{1} - \boldsymbol{\theta}\boldsymbol{\mathcal{B}})^{-1} = \left(\mathbf{1} + \boldsymbol{\theta}\boldsymbol{\mathcal{B}} + \boldsymbol{\theta}^{\mathbf{2}}\boldsymbol{\mathcal{B}}^{\mathbf{2}} + \cdots + \boldsymbol{\theta}^{\boldsymbol{k}}\boldsymbol{\mathcal{B}}^{\boldsymbol{k}}\right)\left(\mathbf{1} - \boldsymbol{\theta}^{\boldsymbol{k}+\mathbf{1}}\boldsymbol{\mathcal{B}}^{\boldsymbol{k}+\mathbf{1}}\right)^{-\mathbf{1}}\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}$$

Ou seja

$$\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}} = -\boldsymbol{\theta}\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}-\mathbf{1}} - \boldsymbol{\theta}^{\mathbf{2}}\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}-\mathbf{2}} - \cdots - \boldsymbol{\theta}^{\boldsymbol{k}}\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}-\boldsymbol{k}} + \boldsymbol{a}_{\boldsymbol{t}} - \boldsymbol{\theta}^{\boldsymbol{k}+\mathbf{1}}\boldsymbol{a}_{\boldsymbol{t}-\boldsymbol{k}+\mathbf{1}} \tag{4.87}$$

Se $|\theta| < 1$, fazendo *k* tender para infinito, obtém-se a série infinita

$$\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}} = -\boldsymbol{\theta}\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}-\mathbf{1}} - \boldsymbol{\theta}^{\mathbf{2}}\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}-\mathbf{2}} - \cdots + \boldsymbol{a}_{\boldsymbol{t}} \tag{4.88}$$

E as ponderações $\boldsymbol{\pi}$ do modelo na forma (4.85) são $\boldsymbol{\pi}_j = -\boldsymbol{\theta}^j$, pelo que, se $|\boldsymbol{\theta}| \geq 1$ os desvios presentes $\tilde{Y}_t$ dependem de $\tilde{Y}_{t-1}, \tilde{Y}_{t-2}, \ldots, \tilde{Y}_{t-k}$ com ponderações que incrementa com k. Para evitar esta situação requer-se que $|\boldsymbol{\theta}| < 1$. Dir-se-á então que a série é invertível. Esta condição é equivalente a que $\sum_{j=0}^{\infty}|\boldsymbol{\theta}|^j \equiv \sum_{j=0}^{\infty}|\boldsymbol{\pi}_j| < \infty$ tal que a série

$$\boldsymbol{\pi}(\boldsymbol{\mathcal{B}}) = (\mathbf{1} - \boldsymbol{\theta}\boldsymbol{\mathcal{B}})^{-\mathbf{1}} = \sum_{\boldsymbol{j}=\mathbf{0}}^{\infty}\boldsymbol{\theta}^{\boldsymbol{j}}\,\boldsymbol{\mathcal{B}}^{\boldsymbol{j}}$$

Converge para todo o $|\mathcal{B}| \leq 1$, ou seja, $\mathcal{B}$ encontra-se dentro ou na fronteira do circulo unitário (Box, Jenkins, & Reinsel, 2008).

A condição de invertibilidade para um processo de ordem superior pode obter-se escrevendo (4.85) na forma

$$\boldsymbol{a}_{\boldsymbol{t}} = \boldsymbol{\theta}^{-\mathbf{1}}(\boldsymbol{\mathcal{B}})\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}$$

Assim, se

$$\boldsymbol{\theta}(\boldsymbol{\mathcal{B}}) = \prod_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{q}}(\mathbf{1} - \boldsymbol{H}_{\boldsymbol{i}}\boldsymbol{\mathcal{B}})$$

onde ${H_1}^{-1}, \ldots, {H_q}^{-1}$ são as raízes de $\theta(\mathcal{B}) = 0$, então, recorrendo ao método de expansão em fracções parciais, obtém-se

$$\boldsymbol{\pi}(\boldsymbol{\mathcal{B}}) = \boldsymbol{\theta}^{-\mathbf{1}}(\boldsymbol{\mathcal{B}}) = \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{q}}\frac{\boldsymbol{M}_{\boldsymbol{i}}}{\mathbf{1} - \boldsymbol{H}_{\boldsymbol{i}}\boldsymbol{\mathcal{B}}}$$

Que converge se $|H_i| < 1$, para $i = 1, 2, \dots, q$. Tem-se então que a condição de invertibilidade para um processo MA(q) é que as raízes H_i^{-1} da equação característica

$$\boldsymbol{\theta(\mathcal{B}) = 1 - \theta_1 \mathcal{B} - \theta_2 \mathcal{B}^2 - \cdots - \theta_q \mathcal{B}^q} \quad \textbf{(4.89)}$$

Caem fora do ciclo unitário (Box, Jenkins, & Reinsel, 2008).

4.2.2.1 Momentos de um processo MA(q)

Desde que um processo MA(q) seja estacionário, a sua média é dada pelo termo constante definido na equação às diferenças **(4.83)**

$$\boldsymbol{E[Y_t] = \mu}$$

A variância obtém-se de forma imediata pela aplicação do operador de variância à equação **(4.83)**

$$\boldsymbol{\gamma_0 \equiv \sigma_Y^2 = Var[\, Y_t] = \left(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2\right)\sigma_\varepsilon^2}$$

A função autocovariância obtém-se a partir de

$$\boldsymbol{\gamma_k = E\left[\left(\varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}\right)\left(\varepsilon_{t-k} - \theta_1 \varepsilon_{t-k-1} - \cdots - \theta_q \varepsilon_{t-k-q}\right)\right]}$$

Apenas os produtos cruzados contemporâneos (para o mesmo instante) dos $\boldsymbol{\varepsilon's}$ são diferentes de zero, pelo que a equação anterior fica na forma

$$\boldsymbol{\gamma_k = \begin{cases} 0 & se\ k > q \\ \left(-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \cdots + \theta_{t-k}\theta_q\right)\sigma_\varepsilon^2 & se\ k = 1, 2, \dots, q \end{cases}}$$

A função autocorrelação surge assim de forma automática através da divisão de $\boldsymbol{\gamma_k}$ por $\boldsymbol{\gamma_0}$

$$\boldsymbol{\rho_k = \begin{cases} 0 & se\ k > q \\ \dfrac{-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \cdots + \theta_{t-k}\theta_q}{1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2} & se\ k = 1, 2, \dots, q \end{cases}} \quad \textbf{(4.90)}$$

Verifica-se que a função de autocorrelação de um processo MA(q) é zero após a ordem q do processo. Por outras palavras, a função de autocorrelação de um processo média móvel corta a partir da lag q.

Se os valores de $\rho_1, \rho_2, \dots, \rho_q$ são conhecidos, as q equações (4.90) podem ser resolvidas para os parâmetros $\theta_1, \theta_2, \dots, \theta_q$, No entanto, ao contrário das equações de Yule-Walker para um processo autoregressivo, estas equações são não lineares, pelo que, excepto para o caso em que q é igual a um, estas equações tem de ser resolvidas iterativamente como descrito no apêndice A6.2 de (Box, Jenkins, & Reinsel, 2008). De igual modo, também, ao contrário das estimativas para o modelo autoregressivo, as estimativas resultantes podem

não ter grande eficiência estatística, no entanto podem fornecer uma estimativa grosseira bastante útil para a identificação do modelo.

Função Autocorrelação Parcial para MA(1)

Substituindo $q = 1$ em (4.90) obtém-se a função autocorrelação para um processo MA(1)

$$\boldsymbol{\rho_k} = \begin{cases} \dfrac{-\boldsymbol{\theta_1}}{\mathbf{1} + \boldsymbol{\theta_1}^{\mathbf{2}}} & \boldsymbol{k = 1} \\ \mathbf{0} & \boldsymbol{k = 0} \end{cases}$$

Substituindo esta equação na equação (4.81), obtém-se após alguma álgebra (Box, Jenkins, & Reinsel, 2008)

$$\boldsymbol{\phi_{kk}} = \frac{-\boldsymbol{\theta_1}^{\boldsymbol{k}}(\mathbf{1} - \boldsymbol{\theta_1})}{\mathbf{1} - \boldsymbol{\theta_1}^{\mathbf{2}(\boldsymbol{k}+\mathbf{1})}}$$

Verifica-se assim que $|\boldsymbol{\phi_{kk}}| < |\boldsymbol{\theta_1}|^{\boldsymbol{k}}$, e que a função autocorrelação parcial é dominada por um decaimento exponencial. Se $\boldsymbol{\rho_1}$ é positivo, então $\boldsymbol{\theta_1}$ é negativo e a função autocorrelação parcial alterna no sinal. Se $\boldsymbol{\rho_1}$ é negativo, então $\boldsymbol{\theta_1}$ é positivo e a função autocorrelação parcial é negativa.

Dualidade entre processos autoregressivos e média móvel

Os resultados anteriores mostram alguns aspectos da dualidade os processos autoregressivos e média novel. Essa dualidade tem algumas consequências resumidas na Tabela 4.1 retirada de (Box, Jenkins, & Reinsel, 2008)

4.2.3 Modelos ARMA

Já se verificou que um processo média móvel *finito*, p.ex. (4.86), pode ser escrito na forma de um processo autoregressivo *infinito* (4.88). Neste caso, se o processo era realmente um MA(1), foi possível obter uma representação "não parcimónia" (Box, Jenkins, & Reinsel, 2008, p. 16) em termos de um modelo autoregressivo. Inversamente, um processo AR(1) não poderia ser representado parcimoniosamente através de um modelo média móvel . Na prática, para obter parametrizações parcimónias, será por vezes necessário incluir no modelo uma combinação de termos autoregressivos e média móvel no mesmo modelo.

Um modelo misto *autoregressivo – média móvel*, ARIMA(*p,0,q*) ou ARMA(*p*,*q*), representado na forma (4.67), depois de retiradas a componente "I" apresenta a forma

$$\left(\mathbf{1} + \boldsymbol{a_1}\boldsymbol{\mathcal{B}} + \cdots + \boldsymbol{a_p}\boldsymbol{\mathcal{B}}^{\boldsymbol{p}}\right)\boldsymbol{y}(\boldsymbol{t}) \;= \left(\mathbf{1} + \boldsymbol{c_1}\boldsymbol{\mathcal{B}} + \cdots + \boldsymbol{c_q}\boldsymbol{\mathcal{B}}^{\boldsymbol{q}}\right)\boldsymbol{\varepsilon}(\boldsymbol{t})$$

Rearranjando os termos e adaptando a notação (Del Castillo, 2002) / (Box, Jenkins, & Reinsel, 2008), tem-se que um processo ARMA(*p,q*) pode ser descrito por

$$\boldsymbol{Y_t} = \boldsymbol{\mu} + \boldsymbol{\phi_1}\boldsymbol{Y_{t-1}} + \cdots + \boldsymbol{\phi_p}\boldsymbol{Y_{t-p}} - \boldsymbol{\theta_1}\boldsymbol{\varepsilon_{t-1}} - \cdots - \boldsymbol{\theta_q}\boldsymbol{\varepsilon_{t-q}} + \boldsymbol{\varepsilon_t}$$

Utilizando novamente a notação $\tilde{Y}_t = Y_t - \mu$, um processo ARMA pode ser descrito na forma

$$\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}} = \boldsymbol{\phi}_1\widetilde{\boldsymbol{Y}}_{\boldsymbol{t-1}} + \cdots + \boldsymbol{\phi}_p\widetilde{\boldsymbol{Y}}_{\boldsymbol{t-p}} - \ \boldsymbol{\theta}_1\boldsymbol{\varepsilon}_{\boldsymbol{t-1}} - \cdots - \boldsymbol{\theta}_q\boldsymbol{\varepsilon}_{\boldsymbol{t-q}} + \boldsymbol{\varepsilon}_{\boldsymbol{t}} \quad \textbf{(4.91)}$$

Que também pode ter a forma

$$\left(\mathbf{1} - \boldsymbol{\phi}_1\boldsymbol{\mathcal{B}} - \boldsymbol{\phi}_2\boldsymbol{\mathcal{B}}^{\mathbf{2}} - \cdots - \boldsymbol{\phi}_p\boldsymbol{\mathcal{B}}^{\boldsymbol{p}}\right)\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}} = \left(\mathbf{1} - \ \boldsymbol{\theta}_1\boldsymbol{\mathcal{B}} - \boldsymbol{\theta}_2\boldsymbol{\mathcal{B}}^{\mathbf{2}} - \cdots - \boldsymbol{\theta}_q\boldsymbol{\mathcal{B}}^{\boldsymbol{q}}\right)\boldsymbol{\varepsilon}_{\boldsymbol{t}}$$

Ou

$$\boldsymbol{A}(\boldsymbol{\mathcal{B}})\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}} = \boldsymbol{C}(\boldsymbol{\mathcal{B}})\boldsymbol{\varepsilon}_{\boldsymbol{t}}$$

Onde $A(\mathcal{B})$ e $C(\mathcal{B})$ são operadores polinomiais em $\mathcal{B}$ de grau *p* e *q* respectivamente.

Um modelo ARMA(*p,q*) é estacionário se a componente AR é estacionária, isto é, se o polinómio $A(\mathcal{B})$ tem todas as raízes fora do circulo unitário no plano $\mathcal{B}$. Similarmente, o processo é invertível se a parte MA é invertível, isto é, se o polinómio $C(\mathcal{B})$ tem todas as raízes fora do circulo unitário.

4.2.3.1 Momentos de um processo ARMA(*p,q*)

Valor esperado
O valor esperado (média) de um processo ARMA $\{Y_t\}$ é dada por (Del Castillo, 2002)

$$\boldsymbol{E}[\boldsymbol{Y}_{\boldsymbol{t}}] = \boldsymbol{\mu} = \frac{\boldsymbol{\xi}}{\mathbf{1} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\phi}_{\boldsymbol{j}}}$$

Autocovariância
A autocovariância obtém-se tirando o valor esperado do produto de **(4.91)** por $\tilde{Y}_{t-k}$ (Del Castillo, 2002), (Box, Jenkins, & Reinsel, 2008)

$$\boldsymbol{\gamma}_{\boldsymbol{k}} \ = \ \boldsymbol{\phi}_1\boldsymbol{\gamma}_{\boldsymbol{k-1}} + \cdots + \boldsymbol{\phi}_p\boldsymbol{\gamma}_{\boldsymbol{k-p}} + \boldsymbol{\gamma}_{\boldsymbol{Y}_\varepsilon}(\boldsymbol{k}) - \boldsymbol{\theta}_1\boldsymbol{\gamma}_{\boldsymbol{Y}_\varepsilon}(\boldsymbol{k-1}) - \cdots - \boldsymbol{\theta}_q\boldsymbol{\gamma}_{\boldsymbol{Y}_\varepsilon}(\boldsymbol{k-q})$$

Onde $\boldsymbol{\gamma}_{\boldsymbol{Y}_\varepsilon}(\boldsymbol{k}) = \boldsymbol{E}\left[\widetilde{\boldsymbol{Y}}_{\boldsymbol{t-k}}\,\boldsymbol{\varepsilon}_{\boldsymbol{t}}\right] = \boldsymbol{E}\left[\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}\,\boldsymbol{\varepsilon}_{\boldsymbol{t-k}}\right]$ é a covariância (cruzada) entre Y e $\boldsymbol{\varepsilon}$, será nula se $k > 0$. Como Y_{t-k} depende apenas dos shocks $(\boldsymbol{\varepsilon})$ que ocorreram acima de $t-k$ numa representação média móvel *infinita* $\widetilde{\boldsymbol{Y}}_{\boldsymbol{t-k}} = \sum_{j=0}^{\infty}\psi_j\,\epsilon_{t-k-j}$, segue-se que

$$\boldsymbol{\gamma}_{\boldsymbol{Y}_\varepsilon}(\boldsymbol{k}) = \begin{cases} \mathbf{0} & \boldsymbol{k} > q \\ \boldsymbol{\psi}_{\boldsymbol{-k}}\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}{}^{\mathbf{2}} & \boldsymbol{k} \leq \mathbf{0} \end{cases}$$

Deste modo a equação anterior pode ser expressa por

$$\boldsymbol{\gamma}_{\boldsymbol{k}} \ = \ \boldsymbol{\phi}_1\boldsymbol{\gamma}_{\boldsymbol{k-1}} + \cdots + \boldsymbol{\phi}_p\boldsymbol{\gamma}_{\boldsymbol{k-p}} - \boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}{}^{\mathbf{2}}\left(\boldsymbol{\theta}_k\boldsymbol{\psi}_{\mathbf{0}} + \boldsymbol{\theta}_{\boldsymbol{k+1}}\boldsymbol{\psi}_{\mathbf{1}} + \cdots + \boldsymbol{\theta}_q\boldsymbol{\psi}_{\boldsymbol{q-k}}\right) \quad \textbf{(4.92)}$$

Com a convenção de que $\theta_0 = -1$. Verifica-se que a equação anterior implica que

$$\boldsymbol{\gamma}_{\boldsymbol{k}} \ = \ \boldsymbol{\phi}_1\boldsymbol{\gamma}_{\boldsymbol{k-1}} + \boldsymbol{\phi}_2\boldsymbol{\gamma}_{\boldsymbol{k-2}} + \cdots + \boldsymbol{\phi}_p\boldsymbol{\gamma}_{\boldsymbol{k-p}} \qquad \boldsymbol{k} \geq \boldsymbol{q} + \mathbf{1}$$

E consequentemente

$$\boldsymbol{\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + \cdots + \phi_p\rho_{k-p}} \qquad \boldsymbol{k \geq q+1} \qquad \textbf{(4.93)}$$

ou

$$\boldsymbol{A(\mathcal{B})\rho_k = 0} \qquad \boldsymbol{k \geq q+1}$$

Verifica-se que **(4.92)** não tem solução única, pelo que, para um processo ARMA(p,q) existirão *q* coeficientes de autocorrelação $\boldsymbol{\rho_q, \rho_{q-1}, \ldots, \rho_1}$ cujos valores dependem directamente da escolha dos q parâmetros média móvel $\boldsymbol{\theta}$, bem como nos *p* parâmetros autoregressivos $\boldsymbol{\phi}$. Além disso, os *p* valores $\boldsymbol{\rho_q, \rho_{q-1}, \ldots, \rho_{p-q+1}}$ fornecem os necessários valores de partida para a equação às diferenças $A(\mathcal{B})\rho_k = 0$, onde $k \geq q+1$, os quais, então determinam inteiramente a autocorrelação das *lags* mais altas.

Variância
Fazendo $k = 0$ tem-se

$$\boldsymbol{\gamma_0 = \phi_1\gamma_1 + \cdots + \phi_p\gamma_p + \sigma_\varepsilon{}^2\left(1 - \theta_1\psi_1 - \cdots - \theta_q\psi_q\right)}$$

que tem que ser resolvida juntamente com as *p* equações **(4.92)** para k= $1, 2, \ldots, p$ para se obter $\gamma_0, \gamma_1, \ldots, \gamma_p$.

Tabela 4.1 Resumo das propriedades dos modelos ARMA

	Processo AR	**Processo MA**	**Processo ARMA**
Modelo em termos de $\tilde{Y}$'s	$A(\mathcal{B})\tilde{Y} = \epsilon_t$	$C^{-1}(\mathcal{B})\tilde{Y} = \epsilon_t$	$C^{-1}(\mathcal{B})A(\mathcal{B})\tilde{Y} = \epsilon_t$
Modelo em termos de $\varepsilon's$	$\tilde{Y} = A^{-1}(\mathcal{B})\epsilon_t$	$\tilde{Y} = C(\mathcal{B})\epsilon_t$	$\tilde{Y} = A^{-1}(\mathcal{B})C(\mathcal{B})\epsilon_t$
Pesos π	Série finita	Série infinita	Série infinita
Pesos ψ	Série infinita	Série finita	Série infinita
Estacionaridade	Raízes de $A(\mathcal{B}) = 0$ fora do círculo unitário	Sempre estacionário	Raízes de $A(\mathcal{B}) = 0$ fora do círculo unitário
Invertibilidade	Sempre invertivel	Raízes de $C(\mathcal{B}) = 0$ fora do círculo unitário	Raízes de $C(\mathcal{B}) = 0$ fora do círculo unitário
ACF	Infinita (decaimento exponencial e/ou decaimento sinosoidal	Finita	Infinita (decaimento exponencial e/ou decaimento sinosoidal
	Acaba em cauda (*Tails off*)	Corta após a *lag q*	Acaba em cauda (*Tails off*)
PACF	Finita	Infinita (dominada por decaimento exponencial e/ou decaimento sinosoidal	Infinita (dominada por decaimento exponencial e/ou decaimento sinosoidal
	Corta após a *lag p*	Acaba em cauda (*Tails off*)	*Acaba em cauda* (*Tails off*)

4.2.4 Processos não estacionários: ARIMA

Muitos processos industriais são não estacionários, no sentido de que, saem por eles próprios, sem interferência externa (controlo), de uma situação de descontrolo, ou seja, o processo vagueia ou oscila fora dos valores pretendidos ou objectivo (*Target*). Contudo, um processo não estacionário ARIMA exibe comportamento "explosivo" para raízes estritamente dentro do ciclo unitário.

(Box, Jenkins, & Reinsel, 2008) propõe o uso de processos autoregressivos (AR) que apesar de não estacionários, no sentido não tem média ou variância constante, não tem comportamento explosivo. Viu-se que um processo ARMA é estacionário se as raízes de $A(\mathcal{B}) = 0$ se situam fora do ciclo unitário, por outro lado, exibem comportamento explosivo se alguma raiz se situa estritamente dentro do ciclo unitário. Existe uma terceira hipótese que é quando as raízes de $A(\mathcal{B}) = 0$ se situam na fronteira, ou seja quando tem exactamente o valor 1. (Box, Jenkins, & Reinsel, 2008) refere-se a estes modelos como modelos não estacionários homogéneos (*homogeneous nonstacionary models*). Estes modelos não tem comportamento explosivo, uma vez que a sua média, ou valor esperado é constante, contudo, momentos de maior ordem podem ser não constantes.

Estes processos autoregressivos tem normalmente *d* raízes iguais a 1, na fronteira do circulo unitário, e as restantes maiores que, portanto fora do circulo unitário.

Considere-se por exemplo o processo

$$\boldsymbol{G(\mathcal{B})\tilde{Y}_t = C(\mathcal{B})\varepsilon_t}$$

Onde $G(\mathcal{B})$ é operador autoregressivo não estacionário com *d* raízes iguais a um e as restantes maiores que um. De acordo com (4.67), este modelo pode ser rescrito na seguinte forma

$$\boldsymbol{G(\mathcal{B})\tilde{Y}_t = A(\mathcal{B})\,(1-\mathcal{B})^d\,\tilde{Y}_t = C(\mathcal{B})\varepsilon_t}$$

Onde $A(\mathcal{B})$ é um operador autoregressivo estacionário. Introduzindo a notação $\nabla^{\mathrm{d}}\,\tilde{Y}_t = \nabla^{\mathrm{d}}\,\boldsymbol{Y}_t$ para $d \geq 1$ onde $\boldsymbol{\nabla = (1-\mathcal{B})}$ é o operador diferencial, o modelo pode ser escrito na forma

$$\boldsymbol{A(\mathcal{B})\,\nabla^{\mathrm{d}}\,\tilde{Y}_t = C(\mathcal{B})\varepsilon_t}$$

De forma equivalente, o processo pode ser definido com recurso a duas equações

$$\begin{cases} \boldsymbol{w_t = \nabla^{\mathrm{d}}\,Y_t} \\ \boldsymbol{A(\mathcal{B})\,w_t = C(\mathcal{B})\varepsilon_t} \end{cases}$$

Deste modo, verifica-se que diferenciando a série d vezes, esta pode ser representada por um modelo ARMA estacionário e invertível.

A forma geral de um processo ARMA não estacionário homogéneo é

$$\boldsymbol{A(\mathcal{B})_p\,(1-\mathcal{B})^d\,Y_t = C(\mathcal{B})_q\varepsilon_t} \quad \textbf{(4.94)}$$

Que normalmente é referido com sendo um processo ARIMA(*p,d,q*). Para analisar um processo ARIMA(*p,d,q*), normalmente começa-se por diferenciar a série sucessivamente até esta aparentar um processo estacionário, determinando-se assim o grau de diferenciação *d*. O modelo obtido após a diferenciação será tratado como um modelo ARMA estacionário.

Expressão geral para as ponderações (Pesos) ψ

Um modelo ARIMA $G(\mathcal{B})\tilde{Y}_t = C(\mathcal{B})\varepsilon_t$ onde $G(\mathcal{B})\tilde{Y}_t = A(\mathcal{B})\,(1-\mathcal{B})^d\,\tilde{Y}_t$ pode ser representado de várias formas. Uma das formas consiste numa soma ponderada infinita dos "*shocks*" $(\boldsymbol{\varepsilon_j})$ passados e presente

$$\begin{aligned} \boldsymbol{Y_t} &= \boldsymbol{\varepsilon_t + \psi_1\varepsilon_{t-1} + \psi_2\varepsilon_{t-2} + \cdots} \\ &= \boldsymbol{\varepsilon_t + \sum_{j=1}^{\infty}\psi_j\varepsilon_{t-j} = \sum_{j=0}^{\infty}\psi_j\varepsilon_{t-j}} \\ &= \boldsymbol{\psi(\mathcal{B})\,\varepsilon_t} \end{aligned} \quad \textbf{(4.95)}$$

Em que $\psi_0 = 1$ e, para simplificação e sem perda de generalidade, se fez $\mu = 0$. Se em ambos os lados da equação (4.95) se aplicar um operador autoregressivo generalizado $\varphi(\mathcal{B})$ dado por

$$\begin{aligned} \boldsymbol{\varphi(\mathcal{B})} &= \boldsymbol{\left(1-\phi_1\mathcal{B}-\cdots-\phi_p\mathcal{B}^p\right)(1-\mathcal{B})^d} \\ &= \boldsymbol{1-\varphi_1\mathcal{B}-\cdots-\varphi_{p+d}\mathcal{B}^{p+d}} \end{aligned} \quad \textbf{(4.96)}$$

Obtém-se

$$\boldsymbol{\varphi(\mathcal{B})Y_t = \varphi(\mathcal{B})\psi(\mathcal{B})\,\varepsilon_t}$$

Comparando o membro direito da equação anterior com o membro direito do modelo geral de um processo ARIMA (4.95), tem-se que

$$\boldsymbol{\varphi(\mathcal{B})Y_t = C(\mathcal{B})\,\varepsilon_t} \quad \textbf{(4.97)}$$

Segue-se então que

$$\boldsymbol{\varphi(\mathcal{B})\psi(\mathcal{B}) = C(\mathcal{B})} \quad \textbf{(4.98)}$$

Deste modo, as ponderações $\boldsymbol{\psi_j}$ podem ser determinadas igualando os coeficientes de $\boldsymbol{\mathcal{B}}$ na expansão da equação anterior, ou seja, de (4.96) tem-se que

$$\underbrace{\left(\mathbf{1}-\boldsymbol{\varphi_1}\mathcal{B}-\cdots-\boldsymbol{\varphi_{p+d}}\mathcal{B}^{p+d}\right)}_{\boldsymbol{\varphi(\mathcal{B})}}\underbrace{\left(\mathbf{1}+\boldsymbol{\psi_1}\mathcal{B}+\boldsymbol{\psi_2}\mathcal{B}^{2}+\cdots\right)}_{\boldsymbol{\psi(\mathcal{B})}}=\underbrace{\left(\mathbf{1}-\boldsymbol{\theta_1}\mathcal{B}-\cdots-\boldsymbol{\theta_1}\mathcal{B}^{q}\right)}_{\boldsymbol{C(\mathcal{B})}} \quad \textbf{(4.99)}$$

Assim, pode conclui-se que as ponderações ψ_j de um processo ARIMA podem ser determinadas recursivamente através das equações (Box, Jenkins, & Reinsel, 2008)

$$\boldsymbol{\psi_j = \varphi_1\psi_{j-1} + \varphi_2\psi_{j-2} + \cdots + \varphi_{p+d}\psi_{j-p+d} - \theta_j} \qquad \boldsymbol{j} > 0$$

Com $\psi_0 = 1$, $\psi_j = 0$ para $j < 0$, e $\theta_j = 0$ para $j > q$.

Forma invertida
No ponto 4.2.2 verificou-se que o modelo

$$\boldsymbol{Y_t = \psi(\mathcal{B})\epsilon_t}$$

Também pode ser escrito de forma invertida

$$\boldsymbol{\psi^{-1}(\mathcal{B})Y_t = \epsilon_t}$$

ou

$$\boldsymbol{\pi(\mathcal{B})Y_t = \left(1-\sum_{j=1}^{\infty}\pi_j\mathcal{B}^j\right)Y_t = \epsilon_t} \quad \textbf{(4.100)}$$

Ou seja, está-se perante uma soma ponderada infinita dos valores observados de z adicionada a um shocks aleatório

$$\boldsymbol{Y_t = \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \cdots + \varepsilon_t}$$

Onde, devido à condição de invertibilidade, os pesos $\boldsymbol{\pi}$ tem de consistir numa série convergente.

Substituindo ϵ_t dado por **(4.100)** na equação **(4.97)**, obtém-se

$$\begin{cases}\boldsymbol{\varphi(\mathcal{B})Y_t = C(\mathcal{B})\,\varepsilon_t}\\ \boldsymbol{\varphi(\mathcal{B})Y_t = C(\mathcal{B})\,\pi(\mathcal{B})Y_t}\end{cases}$$

Ou seja, as ponderações π podem ser determinadas explicitamente igualando os coeficientes de $\boldsymbol{\mathcal{B}}$ na expansão da equação anterior:

$$\boldsymbol{\varphi(\mathcal{B}) = C(\mathcal{B})\,\pi(\mathcal{B})}$$

Isto é:

$$\underbrace{\left(\mathbf{1}-\boldsymbol{\varphi_1}\mathcal{B}-\cdots-\boldsymbol{\varphi_{p+d}}\mathcal{B}^{\boldsymbol{p+d}}\right)}_{\boldsymbol{\varphi(\mathcal{B})}}=\underbrace{\left(\mathbf{1}-\boldsymbol{\theta_1}\mathcal{B}-\cdots-\boldsymbol{\theta_1}\mathcal{B}^{\boldsymbol{q}}\right)}_{\boldsymbol{C(\mathcal{B})}}\underbrace{\left(\mathbf{1}-\boldsymbol{\pi_1}\mathcal{B}-\boldsymbol{\pi_2}\mathcal{B}-\cdots\right)}_{\boldsymbol{\pi(\mathcal{B})}} \quad \textbf{(4.101)}$$

Verifica-se que os pesos π_j de um processo ARIMA podem ser determinados recursivamente através da expressão

$$\boldsymbol{\pi_j}=\boldsymbol{\theta_1\pi_{j-1}}+\boldsymbol{\theta_2\pi_{j-2}}+\cdots+\boldsymbol{\theta_q\pi_{j-q}}+\boldsymbol{\varphi_j} \qquad \boldsymbol{j}>0$$

Com a convenção $\pi_0=1$, $\pi_j=0$ para $\boldsymbol{j}<0$ e $\boldsymbol{\varphi_j}=\mathbf{0}$ para $j>\boldsymbol{p}+\boldsymbol{d}$.

4.2.5 Utilização de modelos ARIMA em previsão

Seja $\hat{Y}_{t+l|t}$ a previsão do valor da variável aleatória Y para o instante $t+l$, Y_{t+l} estimado no instante t. O critério mais comum para calcular estimativas é o critério dos desvios quadráticos médios (*mean square error – MSE*). Se uma previsão for determinada tal que esta minimize o erro quadrático médio, esta será chamada de estimativa da média mínima do quadrado dos desvios (*minimum mean square error – MMSE*) também apelidada de estimativa dos mínimos quadrados. Para determinar a estimativa dos mínimos quadrados de $\hat{Y}_{t+l|t}$ à que determinar

$$\min_{\hat{\boldsymbol{Y}}_{\boldsymbol{t+l|t}}}\boldsymbol{MSE}\left(\hat{\boldsymbol{Y}}_{\boldsymbol{t+l|t}}\right)=\min_{\hat{\boldsymbol{Y}}_{\boldsymbol{t+l|t}}}\boldsymbol{E_t}\left[\left(\boldsymbol{Y_{t+l}}-\hat{\boldsymbol{Y}}_{\boldsymbol{t+l|t}}\right)^{\mathbf{2}}\right]$$

Em que o símbolo $E_t[\cdot]$ significa o valor esperado calculado no instante t. Por outras palavras, qualquer variável aleatória com índice $j>t$, representa um instante situado no futuro, desfasado de j períodos em relação ao instante actual denominado por instante t. Assume-se que no actual instante t, o valor real da variável Y, referente ao período t, já é conhecido.

Uma observação gerada pelo processo no instante $t+l$ poderá ser expressa a partir do modelo (4.95), tomando a forma (Box, Jenkins, & Reinsel, 2008)

$$\boldsymbol{Y_{t+l}}=\sum_{\boldsymbol{j=0}}^{\infty}\boldsymbol{\psi_j\epsilon_{t+l-j}} \quad \textbf{(4.102)}$$

Onde $\psi_0=1$ e, tal como em (4.99), as ponderações ψ podem ser determinadas igualando os coeficientes de $\mathcal{B}$ na igualdade

$$\boldsymbol{\varphi(\mathcal{B})}\left(\mathbf{1}+\boldsymbol{\psi_1}\mathcal{B}+\boldsymbol{\psi_2}\mathcal{B}^{\mathbf{2}}+\cdots\right)=\boldsymbol{C(\mathcal{B})}$$

Equivalentemente, para l positivo, com referência a $k<t$, o modelo pode ser escrito de forma truncada

$$\begin{aligned}\boldsymbol{Y_{t+l}} &= \boldsymbol{\varepsilon_{t+l}}+\boldsymbol{\psi_1\varepsilon_{t+l-1}}+\cdots+\boldsymbol{\psi_{l-1}\varepsilon_{t+1}}+\boldsymbol{\psi_l\varepsilon_t}+\cdots+\boldsymbol{\psi_{t+l-k-1}\varepsilon_{k+1}}+\boldsymbol{C_k(t+l-k)}\\ &= \boldsymbol{\varepsilon_{t+l}}+\boldsymbol{\psi_1\varepsilon_{t+l-1}}+\cdots+\boldsymbol{\psi_{l-1}\varepsilon_{t+1}}+\boldsymbol{C_t(l)}\end{aligned}$$

Onde $C_k(t+l-k)$ é a função complementar relativa à origem finita k do processo

$$\boldsymbol{C_t(l) = \sum_{j=l}^{\infty} \psi_j \epsilon_{t+l-j}}$$

Suponha-se que, na origem t, se pretende fazer uma estimativa $\widehat{\boldsymbol{Y}}_{\boldsymbol{t+l|t}}$ do valor da variável $\boldsymbol{Y_{t+l}}$, que é uma função linear dos valores observados no passado e presente da própria variável, Y_t, Y_{t-1}, Y_{t-2}, ... Então esta será também uma função linear dos shocks $(\boldsymbol{\varepsilon_j})$ passados e presente $\boldsymbol{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}}$, ...

Suponha-se que a melhor estimativa é (Box, Jenkins, & Reinsel, 2008)

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{t+l|t}} = \boldsymbol{\psi_l^* \epsilon_l + \psi_{l+1}^* \epsilon_{l-1} + \psi_{l+2}^* \epsilon_{l-2} + \cdots}$$

Onde os termos ψ_l^*, $\psi_{l+1}^*, \psi_{l+2}^*$, ... são a determinar. Então, usando **(4.102)**, o mínimo "erro médio quadrático, MSE, da estimativa será

$$\boldsymbol{E\left[Y_{t+l} - \widehat{Y}_{t+l|t}\right]^2 = \left(1 + \psi_1^2 + \cdots + \psi_{l+1}^2\right)\sigma_\varepsilon{}^2 + \sum_{j=0}^{\infty}\left(\psi_{l+j} - \psi_{l+j}^*\right)\sigma_\varepsilon{}^2}$$

O qual será minimizado se $\psi_{l+j}^* = \boldsymbol{\psi_{l+j}}$. Tem-se então

$$\boldsymbol{Y_{t+l} = \underbrace{\varepsilon_{t+l} + \psi_1 \varepsilon_{t+l-1} + \cdots + \psi_{l-1}\varepsilon_{t+1}}_{e_t(l)} + \underbrace{\psi_l \varepsilon_t + \psi_{l+1}\varepsilon_{t-1} + \cdots}_{\widehat{Y}_{t+l|t}}}$$

Ou seja

$$\boldsymbol{Y_{t+l} = e_t(l) + \widehat{Y}_{t+l|t}}$$

Onde $\boldsymbol{e_t(l)}$ é o erro da previsão $\widehat{\boldsymbol{Y}}_{\boldsymbol{t+l|t}}$ a uma distância de l períodos. Verifica-se que o que se pode fazer para minimizar o MSE é escolher $\widehat{\boldsymbol{Y}}_{\boldsymbol{t+l|t}}$ igual ao valor esperado $E_t[\boldsymbol{Y_{t+l}}]$, o que dá precisamente a formula geral de uma previsão MMSE (*minimum mean square error*)

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{t+l|t}} = \boldsymbol{E_t[Y_{t+l}]}$$

(Box, Jenkins, & Reinsel, 2008) apresenta várias forma de estimar valores futuros para uma série temporal. Uma forma proposta por (Del Castillo, 2002) para se obter o valor esperado de $\boldsymbol{Y_{t+l}}$ no instante t, $E_t[\boldsymbol{Y_{t+l}}]$, a partir da equação às diferenças, passa por se escrever a respectiva equação para o instante $t+l$ e calcular-se o valor esperado no instante t. Ou seja, na respectiva equação às diferenças

1. Substituir os valores $\boldsymbol{Y_{t-j}}$ $(j \geq 0)$ pelos valores observados

2. Substituir os valores $\boldsymbol{Y_{t+j}}$ $(j \geq 1)$ pelos valores esperados no instante t, $\boldsymbol{\hat{Y}_{t+j|t}}$
3. Substituir os valores de ε_{t-j} $(j \geq 0)$ por $\hat{\varepsilon}_t = \varepsilon_t = Y_{t-j} - \hat{Y}_{t+l|t-j+1}$
4. Substituir os valores de ε_{t+j} $(j \geq 1)$ por $E_t[\varepsilon_{t+j}] = 0$

4.2.6 Identificação de Modelos ARIMA(*p,d,q*)

Na perspectiva de se estar perante um modelo ARIMA, o objectivo em modelação passa por determinar os valores *p*, *d* e *q* e respectivos parâmetros. A metodologia proposta por Box e Jenkins consiste numa abordagem iterativa para a construção do modelo na qual são testadas empiricamente conjecturas sobre o modelo baseadas nos dados da série temporal $\{y_t\}$:

$$Identificação \Leftrightarrow Estimação \Leftrightarrow Teste\ e\ diagnostico \Rightarrow Utilização\ do\ Modelo$$

O primeiro passo para a identificação de um modelo ARIMA a partir de uma dada serie temporal $\{y_t\}$, consiste em determinar de forma aproximada da estrutura do modelo. Ou seja, o principal objectivo nesta fase é obter alguma ideia dos valores *p*, *d* e *q* necessários para estabelecer a estrutura do modelo ARIMA e simultaneamente obter uma estimativa para os parâmetros. O modelo preliminar então obtido servirá de ponto de partida para a aplicação de métodos de estimação mais formais e eficientes.

Nesta etapa pretende-se então identificar uma apropriada subestrutura a partir da estrutura geral de um modelo ARIMA

$$\boldsymbol{A(\mathcal{B})\,\nabla^d\,Y_t = \theta_0 + C(\mathcal{B})\varepsilon_t}$$

Que possa representar uma determinada série temporal. A abordagem será a seguinte:

1. Avaliar a estacionaridade do processo Y_t e, se necessário, diferenciar Y_t tantas vezes quantas as necessárias para se produzir um processo estacionário, desejavelmente reduzindo o processo a um modelo ARMA:

   $$A(\mathcal{B})\,w_t = \theta_0 + C(\mathcal{B})\varepsilon_t$$

   Onde

   $$w_t = (1-\mathcal{B})^d\,Y_t = \nabla^d\,Y_t$$

   O número de vezes que o processo for diferenciado constituirá uma estimativa para o grau de diferenciação *d*.

2. Identificar o modelo ARMA resultante para w_t

A principal ferramenta estatística utilizada nestes dois passos será a função autocorrelação amostral (SACF – *Sample Autocorrelation Function*) e a função autocorrelação parcial amostral (SPACF – *Sample Parcila Autocorrelation Function*). O seu uso permite não só ter uma ideia da estrutura do modelo mas também obter uma estimativa preliminar dos próprios parâmetros.

4.2.6.1 Técnicas de Identificação

Identificação do grau de diferenciação

Na secção Modelos Autoregressivos4.2.3 viu-se que para um modelo estacionário misto autoregressivo - média móvel ARMA(p,q), ou ARIMA(p,0,q), dado pela expressão $A(\mathcal{B})Y_t = C(\mathcal{B})\varepsilon_t$, a função autocorrelação satisfaz a equação às diferenças

$$\boldsymbol{A(\mathcal{B})\,\rho_k = 0} \qquad \boldsymbol{k > q - p}$$

Além disso, se $A(\mathcal{B}) = \sum_{i=1}^{p}(1 - G_i\mathcal{B})$, a solução da equação às diferenças para a *k*-ésima autocorrelação é, assumindo raízes distintas, da forma

$$\boldsymbol{\rho_k = A_1 G_1{}^k + A_2 G_2{}^k + \cdots + A_q G_q{}^k} \qquad \boldsymbol{k > q - p} \qquad \textbf{(4.103)}$$

A condição de estacionaridade requer que os zeros de $\boldsymbol{A(\mathcal{B})}$ caiam fora do circulo unitário, o que implica que as raízes $G_1, G_2, \ldots, G_q$ caiam dentro do circulo unitário.

A equação anterior mostra que nos casos de modelos estacionários em que nenhumas das raízes se situam perto da fronteira do círculo unitário, a função autocorrelação extinguir-se-á rapidamente para valores grandes ou moderados de k. Suponha-se que se está perante o a existência de uma única raiz real próxima de um, tal que $G_1 = 1 - \delta$ onde δ representa uma quantidade bastante pequena. Então, para k grande tem-se que $\rho_k \simeq A_1(1 - k\delta)$ e a função autocorrelação decairá lentamente e muito próximo da linearidade. O mesmo argumento pode ser aplicado se mais do que uma raiz se aproxima da unidade. Assim, a tendência para a função autocorrelação não se extinguir é tida como uma indicação da possível existência de uma raiz próxima da unidade. A função autocorrelação estimada tende a seguir o comportamento da função autocorrelação teórica (Box, Jenkins, & Reinsel, 2008). Nesses casos deve-se estar perante processos estocásticos não estacionários, mas em que, possivelmente, o processo $w_t = \nabla y_t$ será estacionário.

Normalmente, através de um simples *plot* da série de dados $\{y_t\}$, consegue-se verificar quando é que o processo é estacionário ou não. Se o processo apresenta indícios de não estacionaridade, procede-se à diferenciação da série, obtendo-se uma nova série $w_t = \nabla y_t$ = $y_t - y_{t-1}$. Se a série resultante continuar a apresentar indícios de não estacionaridade, esta será diferenciada novamente até se obter uma série estacionária, determinando-se deste modo o grau de diferenciação *d*. Uma vez obtido um processo estacionário por diferenciações sucessivas, procede-se à identificação de p e q, e à estimativa dos parâmetros do modelo.

Por vezes pode existir o perigo da diferenciação excessiva introduzindo-se assim um termo "média móvel" não invertível. Adicionalmente a variância do modelo sobre diferenciado será inflacionado artificialmente, e isso pode desvirtuar a actual estrutura dos dados. Para ilustra o problema, considere-se por exemplo o seguinte modelo

$$\boldsymbol{Y_t = Y_{t-1} + \varepsilon_t}$$

Diferenciando-se este processo obtém-se um processo constituído por ruído branco ε_t, mas, se por alguma razão se decidir diferenciar novamente, obtém-se

$$\boldsymbol{\nabla^2 Y_t = \nabla \varepsilon_t = (1 - \mathcal{B})\varepsilon_t}$$

Ou seja introduziu-se um termo MA não invertível com a consequência lógica de que será dado mais peso às observações passadas do que às actuais. Consequentemente, a variância da série sobre diferenciada será inflacionada. Neste exemplo, a variância de $\nabla^2 Y_t$ é o dobro da variância de ∇Y_t.

Tem sido proposta várias alternativas para detectar a diferenciação excessiva e assim determinar o valor correcto de *d*. Uma simples abordagem consiste em calcular sucessivamente $Var(y_t)$, $Var(\nabla y_t)$, … , $Var(\nabla^n y_t)$ e escolher *d* tal que a variância seja mínima.

Identificação do processo ARMA resultante

Após a obtenção do grau de diferenciação *d*. o próximo passo passa pelo estudo da aparência das funções de autocorrelação e autocorrelação parcial da série diferenciada $w_t = \nabla^d Y_t$, para se procurar pistas sobre a escolha das ordens *p* e *q* para os operadores autoregressivo e média móvel.

Como se viu nas secções anteriores, enquanto que a função autocorrelação de um processo autoregressivo de ordem *p* acaba em cauda (decaimento exponencial), a sua função autocorrelação parcial corta após a *lag p*. Contrariamente, a função autocorrelação de um processo média móvel de ordem *q* corta após a *lag q*, enquanto a sua função autocorrelação parcial acaba em cauda. Se ambas as funções PACF e ACF acabam em cauda, provavelmente está-se perante um processo misto.

A Tabela 4.2, extraída de (Del Castillo, 2002), fornece alguns elementos úteis para identificação de processos autoregressivos e média móvel

Tabela 4.2 Resultados uteis para identificação de processos AR e MA

	Processo MA(p)	**Processo AR(q)**	**Ruido branco**
ACF$(\boldsymbol{\rho_k})$	$\rho_k = 0$ para $k > q$	ρ_k acaba em cauda (*tails off*)	$\rho_k = 0$ para todo o k
	$\hat{\sigma}(r_k) = \sqrt{\frac{1}{N}\left(1 + 2\sum_{v=1}^{q} r_v^2\right)}$		$\boldsymbol{\hat{\sigma}(r_k) = \frac{1}{N}}$
	Para $k > q$		Para todo o k
PACF$(\boldsymbol{\phi_{kk}})$	$\boldsymbol{\phi_{kk}}$ acaba em cauda (*tails off*)	$(\boldsymbol{\phi_{kk}}) = \mathbf{0}$ para $k > p$	$(\boldsymbol{\phi_{kk}}) = \mathbf{0}$ para todo o k
		$\boldsymbol{\hat{\sigma}(\hat{\phi}_{kk}) = \frac{1}{\sqrt{N}}}$	$\boldsymbol{\hat{\sigma}(\hat{\phi}_{kk}) = \frac{1}{\sqrt{N}}}$
		para $k > p$	Para todo o k

A função auto correlação de um processo misto, contendo uma componente autoregressiva de ordem *p* e uma componente média móvel de ordem *q* é uma mistura de decaimento

exponencial e ondas sinusoidais amortecidas após as primeiras (*q* – *p*) *lags*. Em sentido inverso, a função auto correlação parcial de um processo misto é dominada por uma mistura de decaimento exponencial e ondas sinusoidais amortecidas após as primeiras (*p* – *q*) *lags*

Em geral, o comportamento da função de autocorrelação de um modelo autoregressivo (média móvel) tende a imitar o comportamento da função autocorrelação parcial de um modelo média móvel (autoregressivo) (Box, Jenkins, & Reinsel, 2008).

Desvio padrão para estimativas de autocorrelação e autocorrelação parcial

Quando se faz uma estimativa é sempre importante saber qual o nível de significância dessa estimativa. É sempre importante ter algum indicador que nos dê alguma informação sobre a coerência da estimativa. Neste caso particular, necessita-se de alguma informação que nos permita decidir quando é que a autocorrelação e autocorrelação parcial é efectivamente zero após determinada *lag p* ou *q*. Para um número de *lag* elevado, na hipótese de que se está perante um processo média móvel de ordem *q*, pode-se determinar o desvio padrão da autocorrelação estimada a partir da forma simplificada da fórmula de Bartlett (3.1), substituindo os valores teóricos pelos valores estimados

$$\widehat{\boldsymbol{\sigma}}[\boldsymbol{r_k}] = \frac{\mathbf{1}}{\sqrt{\boldsymbol{N}}}\left[\mathbf{1} + \mathbf{2}\left(\boldsymbol{r_1^2} + \boldsymbol{r_2^2} + \cdots + \boldsymbol{r_{2q}^2}\right)\right]^{\mathbf{1/2}} \qquad \boldsymbol{k} > q$$

Para a autocorrelação parcial utiliza-se o resultado (4.82) que, na hipótese de que se trata de um processo autoregressivo de ordem *p*, o desvio padrão para a autocorrelação parcial de ordem *p+1* e superior é (Box, Jenkins, & Reinsel, 2008)

$$\widehat{\boldsymbol{\sigma}}\left[\widehat{\boldsymbol{\phi}}_{\boldsymbol{kk}}\right] = \frac{\mathbf{1}}{\sqrt{\boldsymbol{N}}} \qquad \boldsymbol{k} > q \qquad \textbf{(4.104)}$$

4.2.6.2 Ferramentas adicionais de identificação

Na secção anterior verificou-se que as funções de autocorrelação e autocorrelação parcial são ferramentas extremamente úteis para identificação de modelos ARIMA a partir de séries temporais, no entanto poderão existirem casos envolvendo modelos mistos em que esta ferramentas apresentem resultados ambíguos. Uma das ferramentas propostas por (Box, Jenkins, & Reinsel, 2008), entre outras, é a análise da correlação canónica.

Método das correlações canónicas

Em geral, para dois conjuntos de variáveis, $Y_1 = (y_{11},\ y_{12},\ \ldots\ , y_{1k})'$ e $Y_2 = (y_{21},\ y_{22},\ \ldots\ , y_{2l})'$, de dimensões *k* e *l* respectivamente (assume-se que $k > l$), a análise de correlação canónica consiste em determinar combinações lineares $U_i = a_i' Y_1$ e $V_i = b_i' Y_2,\ i = 1,\ \ldots\ , k$ e as correspondentes correlações $\rho_i = corr[U_i,\ V_i]$ com $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_k \geq 0$, tal que tal que os U_i e os V_j estão mutuamente não correlacionados para $i \neq j$. U_1 e V_1 tem a máxima correlação possível ρ_1 entre todas as combinações lineares de Y_1 e Y_2, U_2 e V_2 tem a máxima correlação possível ρ_2 entre todas as combinações lineares de Y_1 e Y_2 que não estão correlacionadas com U_1 e V_1, e assim sucessivamente. Às correlações resultantes ρ_i dá-se o nome de correlações canónicas entre Y_1 e Y_2, e as variáveis U_i e V_i são as correspondentes variáveis canónicas. Se $\Omega = cov[Y]$ for a matriz de covariância de $Y = (Y_1', Y_2')'$, com $\Omega_{ij} = cov[Y_i,\ Y_j]$, então é sabido que os valores $\rho^2(i)$ são os valores próprios

ordenados da matriz $\Omega_{11}^{-1}\Omega_{12}\Omega_{22}^{-1}\Omega_{21}$ e os vectores a_i, tal que $U_i = a_i^{'}Y_i$, são os correspondentes vectores próprios normalizados; ou seja, os $\rho^2(i)$ e os a_i satisfazem a condição

$$\left[\boldsymbol{\rho^2(i)I - \Omega_{11}^{-1}\Omega_{12}\Omega_{22}^{-1}\Omega_{21}}\right]\boldsymbol{a_i} = \mathbf{0} \qquad \boldsymbol{i} = \mathbf{1}, \dots, \boldsymbol{k} \tag{4.105}$$

Com $\rho^2(1) \geq \rho^2(12) \geq \cdots \geq \rho^2(k) \geq 0$. De forma similar, pode-se definir a noção de "Correlação Canónica Parcial" entre Y_1 e Y_2, dado outro conjunto de variáveis Y_3 como a correlação canónica entre Y_1 e Y_2, após estas terem sido ajustadas para o efeito de Y_3 por regressão linear sobre Y_3. Especificamente, seja $\Omega_{\text{ij.m}} = \Omega_{\text{ij}} - \Omega_{\text{im}}\Omega_{\text{mm}}^{-1}\Omega_{\text{mj}} = cov\left[Y_{i.m},\ Y_{j.m}\right]$ para $i, j = 1, 2$ e $m = 3$, onde $Y_{i.m} = Y_i - \Omega_{\text{im}}\Omega_{\text{mm}}^{-1}Y_m$ é Y_i ajustado por regressão linear sobre Y_m. Assim, $\Omega_{\text{ij.m}}$ representa a matriz de covariância entre Y_i e Y_j após ajustamento destas variáveis para os efeitos de linearidade de outro conjunto de variáveis Y_m

No contexto de modelos de séries temporais ARMA, esta ferramenta poderá ser de extrema utilidade seguindo como documenta (Box, Jenkins, & Reinsel, 2008) e (Reinsel, 1997). Neste trabalho apenas se aborda dois casos especiais destas correlações canónicas:

1. Primeiro, quando $m = 0$, simplesmente se examina a correlação canónica (autocorrelação ρ_{j+1}) entre Y_t e Y_{t-i-1}, e esta será sempre igual a zero num processo MA(q) para $j \geq q$.
2. Segundo, quando j=0, examina-se a correlação canónica parcial (autocorrelação parcial $\phi_{m+1,m+1}$) entre Y_t e Y_{t-m-1}, dado $Y_{t-1}, \dots, Y_{t-m}$, e esta será sempre igual a zero num processo AR(p) para $m \geq p$.

Deste modo, a análise das correlações canónicas pode ser vista como uma extensão da análise das funções de autocorrelação e autocorrelação parcial do processo.

Na prática, baseado em (4.105), está-se a considerar a correlações canónicas amostrais $\hat{\rho}(i)$, que são determinadas a partir dos valores próprios da matriz

$$\begin{aligned}\left(\sum_{\boldsymbol{t}} \boldsymbol{Y_{m,t}\, Y_{m,t}^{'}}\right)^{-1}\left(\sum_{\boldsymbol{t}} \boldsymbol{Y_{m,t}\, Y_{m,t-j-1}^{'}}\right) \\ \times\left(\sum_{\boldsymbol{t}} \boldsymbol{Y_{m,t-j-1}\, Y_{m,t-j-1}^{'}}\right)^{-1}\left(\sum_{\boldsymbol{t}} \boldsymbol{Y_{m,t-j-1}\, Y_{m,t-j-1}^{'}}\right)\end{aligned} \tag{4.106}$$

Para vários valores de *lag* $j = 0, 1, \dots$ e $m = 0, 1, \dots$

Critério de selecção de modelo

Outra abordagem para a selecção do modelo é o uso de critérios tais como AIC proposto por (Akaike, 1974) ou o *Bayesian Information Criteria* (BIC) proposto por (Schwarz, 1978). Na implementação desta abordagem, uma gama de potenciais modelos ARMA são estimados por métodos de máxima verosimilhança, e cada um dos modelos é avaliado com recurso a critérios tais como AIC (normalizado pela dimensão da amostra *n*) dado por

$$AIC_{p,q} = \frac{-2\ln(verosimilhança\ maximizada) + 2r}{n} \approx \ln(\hat{\sigma}_\varepsilon^2) + r\frac{2}{n} + constante$$

Ou pelo critério relacionado

$$\boldsymbol{BIC_{p,q} = \ln\left(\hat{\sigma}_{\varepsilon}^{2}\right) + r\frac{\ln(n)}{n}}$$

onde $\hat{\sigma}_{\varepsilon}^{2}$ é um estimador de máxima verosimilhança de σ_{ϵ}^{2}, e $r = p + q + 1$ é o numero de parâmetros estimados no modelo, incluindo os temos constantes. Nestes critérios, o segundo termo corresponde a um factor penalizador pela inclusão de parâmetros adicionais no modelo. Nesta abordagem, são preferidos os modelos que resultem num menor valor do critério, em que os valores de AIC e BIC são comparados entre os vários modelos como base de selecção do modelo. Como o critério BIC impõe uma maior penalização para o numero de parâmetros estimados que o AIC, o uso do critério BIC para selecção do modelo deve conduzir a modelo cujo numero de parâmetros não deve ser superior do que o seleccionado pelo critério AIC.

Uma desvantagem desta abordagem é que vários modelos podem ter que ser estimados por máxima verosimilhança, os quais são bastante pesados em termos informáticos, embora essa questão se coloque cada vez menos. Por essa razão, (Hannan & Rissanen, 1982) propôs o seguinte procedimento de selecção de modelo

1. Num primeiro estágio do procedimento obtém-se uma estimativa da série de "shocks" aleatórios ε_t pela aproximação do modelo ARMA desconhecido por um modelo AR (de ordem suficiente alta) de ordem m^*. Esta ordem m^* pode também ela ser escolhida com recurso ao critério AIC. A partir do modelo AR(m^*) obtém-se os resíduos $\tilde{\varepsilon}_t = \tilde{\omega}_t - \sum_{j=1}^{m^*} \hat{\phi}_{m^* j}\, \tilde{\omega}_{t-j}$
2. Num segundo estágio estima-se os parâmetros ϕ_j e θ_j através de regressão linear de $\tilde{\omega}_t$ em $\tilde{\omega}_{t-1}, \ldots, \tilde{\omega}_{t-p}$ e $\tilde{\varepsilon}_{t-1}, \ldots, \tilde{\varepsilon}_{t-q}$, para vários valores de *p* e *q*. Isto é, estima-se modelos aproximados da forma (4.107) com recurso a métodos de regressão dos mínimos quadrados ordinários, e estimar a variância dos erros $\hat{\sigma}_{p,q}^{2}$.

$$\tilde{\omega}_t = \sum_{j=1}^{p} \phi_j\, \tilde{\omega}_{t-j} - \sum_{j=1}^{q} \theta_j\, \tilde{\varepsilon}_{t-j} + \varepsilon_t \qquad \textbf{(4.107)}$$

3. Pela aplicação do critério AIC ou BIC escolhe-se a ordem (p,q) do modelo ARMA que minimize $\ln\left(\hat{\sigma}_{p,q}^{2}\right) + (p+q)\ln(n)/n$

4.2.6.3 Abordagem Box-Jenkins para a modelação de um processo ARIMA

O ajustamento de um modelo a uma série temporal é um processo iterativo. A Figura 4-6, adaptada de (Del Castillo, 2002), mostra a abordagem geral proposta por Box-Jenkins. Se um hipotético modelo ARIMA não demonstrar um bom ajustamento, ou hipóteses alternativas parecerem mais viáveis, estas deverão ser tentadas de forma iterativa. Esta iteração está indicada através da linha tracejada.

Alternativamente, se os resíduos continuarem a revelar alguma autocorrelação em vez de conterem apenas ruído branco, essa deve ser retirada com recurso a um novo modelo ARIMA para os resíduos, ou seja, separar a dinâmica dos resíduos da dinâmica do processo. Esta dinâmica será posteriormente integrada no modelo. Esta iteração está indicada através da linha contínua "paralela" à linha tracejada.

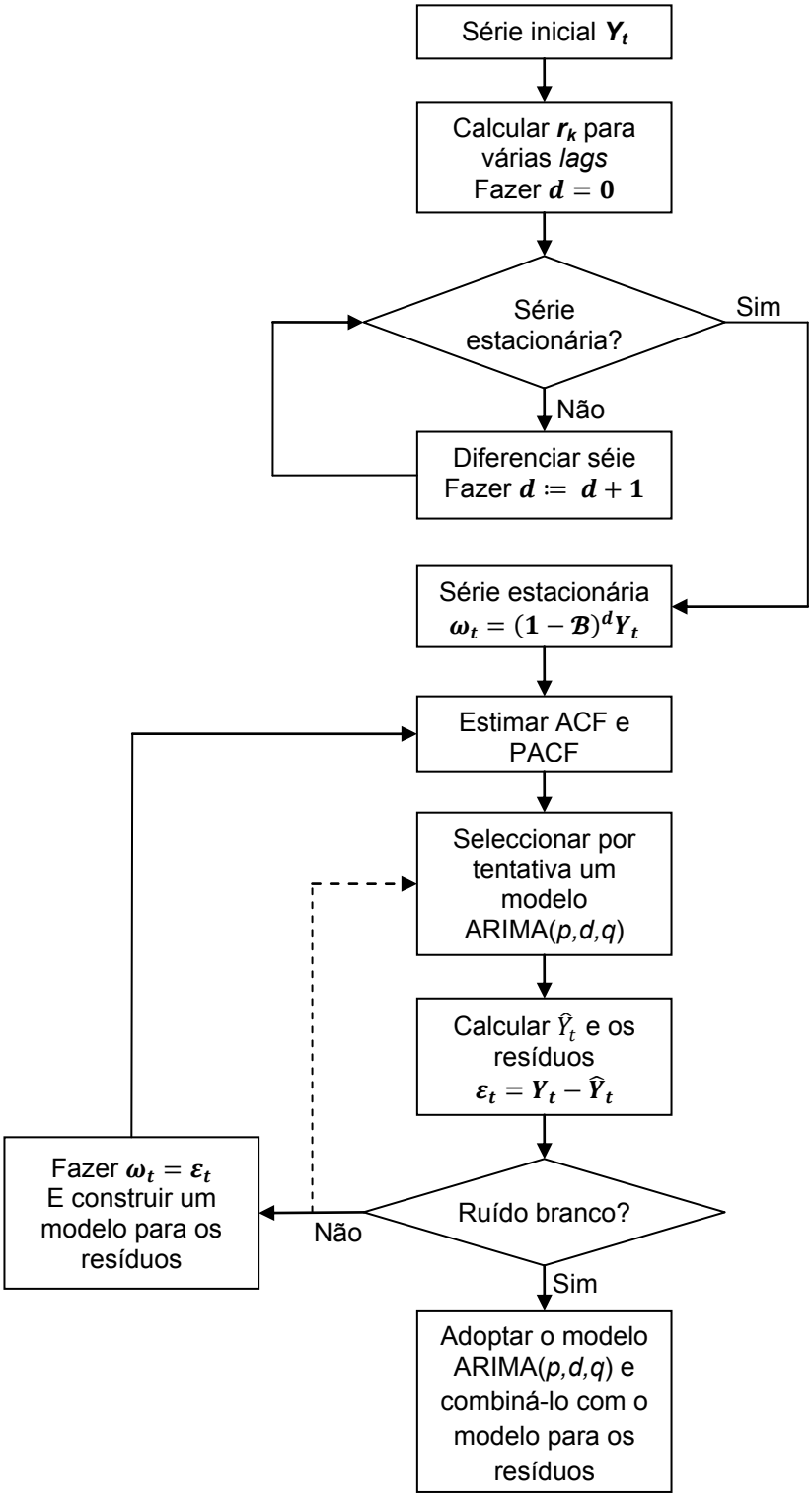


Figura 4-6 Metodologia Box-Jenkins para identificação de processos ARIMA(p,d,q)

4.2.7 Estimação dos Parâmetros do Modelo

Uma vez determinada a estrutura do modelo, o próximo passo será encontrar estimativas eficientes para os parâmetros do modelo. Após a estimação dos parâmetros, o modelo ajustado deve ser sujeito a diagnostico de comprovação e testes da qualidade do ajustamento. Nesta secção apenas se aborda métodos de estimação *offline*. Quer isto dizer

que a estimativa dos parâmetros obtém-se numa só vez após um conjunto de N observações serem colectadas. Isto em contraste com os métodos de estimação *online*, que apesar de não fazerem parte deste trabalho, serão abordados de forma ligeira mais à frente.

4.2.7.1 Estimativa dos mínimos quadrados e máxima verosimilhança

Após a obtenção de uma série temporal $\{y_t\}$ contendo N observações, geradas por um processo ARIMA, e determinada a estrutura do modelo, a etapa seguinte passa por determinar a melhor estimativa dos parâmetros do modelo. Após a diferenciação *d* vezes da série gerada, obtém-se uma nova série estacionária $\{w_t\}$ de dimensão $n = N - d$, onde cada elemento w_t será dado por $w_t = \nabla^{\mathrm{d}} y_t$. O objectivo nesta secção é determinar, a partir da nova série, os parâmetros ϕ_i e θ_i (na notação de (Del Castillo, 2002) entre outros) do modelo ARMA(p,q) resultante:

$$\tilde{\boldsymbol{w}}_{\boldsymbol{t}} - \boldsymbol{\phi}_{\boldsymbol{1}}\tilde{\boldsymbol{w}}_{\boldsymbol{t-1}} - \cdots - \boldsymbol{\phi}_{\boldsymbol{p}}\tilde{\boldsymbol{w}}_{\boldsymbol{t-p}} = \boldsymbol{\theta}_{\boldsymbol{1}}\boldsymbol{\varepsilon}_{\boldsymbol{t-1}} - \boldsymbol{\theta}_{\boldsymbol{2}}\boldsymbol{\varepsilon}_{\boldsymbol{t-2}} - \cdots - \boldsymbol{\theta}_{\boldsymbol{q}}\boldsymbol{\varepsilon}_{\boldsymbol{t-q}} + \boldsymbol{\varepsilon}_{\boldsymbol{t}}$$

Que também pode ser escrito na forma

$$\boldsymbol{\varepsilon}_{\boldsymbol{t}} = \tilde{\boldsymbol{w}}_{\boldsymbol{t}} - \boldsymbol{\phi}_{\boldsymbol{1}}\tilde{\boldsymbol{w}}_{\boldsymbol{t-1}} - \cdots - \boldsymbol{\phi}_{\boldsymbol{p}}\tilde{\boldsymbol{w}}_{\boldsymbol{t-p}} + \boldsymbol{\theta}_{\boldsymbol{1}}\boldsymbol{\varepsilon}_{\boldsymbol{t-1}} + \boldsymbol{\theta}_{\boldsymbol{2}}\boldsymbol{\varepsilon}_{\boldsymbol{t-2}} + \cdots + \boldsymbol{\theta}_{\boldsymbol{q}}\boldsymbol{\varepsilon}_{\boldsymbol{t-q}} \qquad \textbf{(4.108)}$$

Onde $\tilde{\boldsymbol{w}} = \boldsymbol{w}_{\boldsymbol{t}} - \boldsymbol{\mu}$ e $\boldsymbol{\mu} = \boldsymbol{E}[\boldsymbol{w}_{\boldsymbol{t}}]$. Para $d > 0$ poderá por vezes ser apropriado assumir $\boldsymbol{\mu} = 0$, ver discussão em (Box, Jenkins, & Reinsel, 2008) . Quando não for apropriado assumir $\boldsymbol{\mu} = 0$, então assume-se que $\boldsymbol{\mu} \cong \bar{\boldsymbol{w}} = \sum_{t=1}^{n} w_t/n$. Para dimensões amostrais normalmente consideradas na análise de séries temporais, esta aproximação será adequada. No entanto, $\boldsymbol{\mu}$ poderá ser incluído como um parâmetro adicional a ser estimado.

Os valores $w_t's$ não podem ser substituídos directamente na equação anterior para calcular os $\varepsilon_t's$ devido ao problema dos valores iniciais para iniciar a equação às diferenças. Suponha-se que os *p* valores w_* dos $w_t's$ e que os *q* valores ε_* anteriores ao inicio da série são eram dados, então os valores $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ poderiam ser calculados com recurso à equação **(4.108)** condicionados à escolha dos valores iniciais.

Deste modo, para qualquer escolha de parâmetros(ϕ_i, θ_i) e dos valores iniciais (w_*, ε_*), pode-se sempre calcular sucessivamente um conjunto de valores $\varepsilon_t(\phi, \theta | w_*, \varepsilon_*, \mathrm{w})$, $t = 1, 2, \ldots, n$.

Assumindo-se que os $\varepsilon_t's$ são normalmente distribuídos, o que normalmente é um dos pressupostos para a validação do modelo, então a sua densidade probabilidade será pela função verosimilhança

$$\boldsymbol{p}(\boldsymbol{\varepsilon}_{\boldsymbol{1}}, \boldsymbol{\varepsilon}_{\boldsymbol{2}}, \ldots, \boldsymbol{\varepsilon}_{\boldsymbol{n}}) \propto \left(\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\boldsymbol{2}}\right)^{-\boldsymbol{n}/\boldsymbol{2}} \boldsymbol{exp}\left[-\left(\sum_{\boldsymbol{t}=\boldsymbol{1}}^{\boldsymbol{n}} \frac{\boldsymbol{\varepsilon}_{\boldsymbol{t}}^{\boldsymbol{2}}}{\boldsymbol{2}\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\boldsymbol{2}}}\right)\right] = \frac{\boldsymbol{1}}{\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\boldsymbol{n}}} \boldsymbol{exp}\left(\frac{\boldsymbol{SS}(\boldsymbol{\phi}, \boldsymbol{\theta})}{\boldsymbol{2}\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\boldsymbol{2}}}\right)$$

Onde $\phi' = (\phi_1, \phi_2, \ldots, \phi_p)$, $\theta' = (\theta_1, \theta_2, \ldots, \theta_q)$ e $SS(\phi, \theta)$ traduz o somatório dos erros quadráticos como função dos parâmetros (ϕ, θ). Maximizando-se a função verosimilhança com respeito aos parâmetros encontrar-se-á os parâmetros tão compatíveis quanto possível com a evidência da amostra. Neste contexto, o símbolo $\propto$ tem o significado de "ser

proporcional a", onde a constante de proporcionalidade é desconhecida. Tomando o logaritmo da equação anterior, obtém-se a função "verosimilhança logaritmica" (*log-likelihood*) que terá a forma

$$\boldsymbol{L(\phi,\theta,\sigma_\varepsilon) = f(\phi,\theta) - n\ln(\sigma_\varepsilon) - \left(\frac{SS(\phi,\theta)}{2\sigma_\varepsilon^2}\right)}$$

Onde $\boldsymbol{f(\phi,\theta)}$ é uma função de $\boldsymbol{\phi}$ e $\boldsymbol{\theta}$**,** sem interesse nesta altura. Os parâmetros que maximizem esta função (verosimilhança logaritmica, ou *log-likelihood)* são os mesmos que maximizam a função de verosimilhança, pelo que se torna mais simples trabalhar com ela para obter os parâmetros pretendidos.

Os erros são estimados a partir dos resíduos obtidos com recurso à equação **(4.108)** a partir de escolhas dadas de $\boldsymbol{\phi}$ e $\boldsymbol{\theta}$. A soma dos quadrados dos erros (resíduos) é

$$\boldsymbol{SS(\phi,\theta) = \sum_{t=1}^{n} \varepsilon_t^2(\phi,\theta|w_*,\varepsilon_*,\mathrm{w})}$$

A esta forma de cálculo chama-se "soma dos quadrados condicional" porque os seus valores dependem da escolha dos valores iniciais como comentado acima. Para se obter estimativas da máxima verosimilhança (MLE – Maximum Likelihood Estimate) de $\boldsymbol{\phi}$ e $\boldsymbol{\theta}$ maximiza-se $L(\phi,\theta,\sigma_\varepsilon)$. Para amostras de grande dimensão, o termo $\boldsymbol{f(\phi,\theta)}$ torna-se geralmente desprezível, pelo que a maximização de $L(\phi,\theta,\sigma_\varepsilon)$ implica apenas a minimização da soma dos quadrados $SS(\phi,\theta)$. Deste modo, para *n* grande tem-se que a estimativa da máxima verosimilhança (MLE) identifica-se com a estimativa dos mínimos quadrados (OLS – *Ordinary Least Square*), ou seja

$$\boldsymbol{(\hat{\phi},\hat{\theta})_{MLE} = (\hat{\phi},\hat{\theta})_{OLs}}$$

Se o modelo for "um modelo estatístico linear", a estimativa dos mínimos quadrados pode-se obter através de uma fórmula fechada. Os modelos ARIMA são modelos estocásticos, modelos de equações às diferenças mas nem sempre são modelos lineares nos parâmetros. Para os modelos ARIMA, um modelo é linear nos parâmetros se as derivadas dos resíduos, ou seja, as derivadas da equação **(4.108)** em ordem aos parâmetros $\phi_i's$ e $\theta_i's$, não são função de qualquer parâmetro desconhecido.

Estimação de modelos AR(*p*)

É fácil verificar que todos os modelos do autoregressivos puros AR(*p*) são lineares. Um modelo AR(p) pode ser escrito na forma

$$\boldsymbol{\tilde{Y}_t = \phi_1\tilde{Y}_{t-1} + \phi_2\tilde{Y}_{t-2} + \cdots + \phi_p\tilde{Y}_{t-p} + \varepsilon_t}$$

Pelo que os resíduos a serão da forma

$$\boldsymbol{\varepsilon_t = \widetilde{Y}_t - \phi_1 \widetilde{Y}_{t-1} - \phi_2 \widetilde{Y}_{t-2} - \cdots - \phi_p \widetilde{Y}_{t-p}}$$

As derivadas em ordem aos parâmetros serão

$$\boldsymbol{\frac{\partial \varepsilon_t}{\partial \phi_i} = \widetilde{Y}_{t-i} \qquad i = 1, 2, \ldots, p}$$

Que, portanto, será um modelo estatístico linear que poderá ser ajustado com recurso às normais técnicas de regressão linear. Para amostras de grande dimensão, os valores iniciais vão ter um efeito muito reduzido, pelo que uma sequência de variáveis aleatórias definidas por um processo AR(p) pode ser escrito na forma

$$\begin{bmatrix} \widetilde{Y}_{p+1} \\ \widetilde{Y}_{p+2} \\ \vdots \\ \widetilde{Y}_{n-1} \\ \widetilde{Y}_n \end{bmatrix} = \begin{bmatrix} \widetilde{Y}_p & \cdots & \widetilde{Y}_1 \\ \widetilde{Y}_{p+1} & \cdots & \widetilde{Y}_2 \\ \vdots & \vdots & \vdots \\ \widetilde{Y}_{n-2} & \cdots & \widetilde{Y}_{n-p-1} \\ \widetilde{Y}_{n-1} & \cdots & \widetilde{Y}_{n-p} \end{bmatrix} \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_p \end{bmatrix} + \begin{bmatrix} \varepsilon_{p+1} \\ \varepsilon_{p+2} \\ \vdots \\ \varepsilon_{n-1} \\ \varepsilon_n \end{bmatrix}$$

Ou, em notação vectorial

$$\boldsymbol{\widetilde{Y} = Z' \Phi + \varepsilon}$$

Portanto, o estimador dos mínimos quadrados para os parâmetros será

$$\boldsymbol{\hat{\Phi} = (Z' Z)^{-1} Z' \widetilde{Y}} \qquad \textbf{(4.109)}$$

Estimação de modelos mistos

Os modelos ARIMA(*p,d,q*) onde $q > 0$ não podem ser estimados pela minimos quadeados ordinários (formula anterior) porque são não lineares nos parâmetros. (Del Castillo, 2002) demonstra este facto com recurso ao exemplo mais simples, ou seja, um processo MA(1)

$$\boldsymbol{\widetilde{Y}_t = -\theta_1 \varepsilon_{t-1} + \varepsilon_t = (1 - \theta \mathcal{B}) \varepsilon_t}$$

Resolvendo para ε_t e assumindo invertibilidade $(|\boldsymbol{\theta}| < 1)$, tem-se

$$\boldsymbol{\varepsilon_t = \frac{\widetilde{Y}_t}{1 - \theta \mathcal{B}}}$$

Derivado em ordem ao parâmetro

$$\boldsymbol{\frac{\partial \varepsilon_t}{\partial \theta} = \frac{\mathcal{B} \widetilde{Y}_t}{(1 - \theta \mathcal{B})^2} = \frac{\mathcal{B} \varepsilon_t}{1 - \theta \mathcal{B}}}$$

Que é função do parâmetro desconhecido θ, pelo que não será um modelo estatístico linear e, consequentemente, não existirá uma formula fechada para um estimador dos mínimos

quadrados. A solução passará pela obtenção com recurso ao calculo numérico da minimização da função da soma dos quadrados.

(Box, Jenkins, & Reinsel, 2008) propõe um algoritmo iterativo a partir do modelo ARMA, em que para vários valores de ϕ_i e θ_j, igualmente espaçados compreendidos no intervalo $]-1,1[$, se determine os respectivos valores da soma dos quadrados dos resíduos $SS(\phi,\theta)$ obtidos por **(4.108)**, seleccionando-se posteriormente a combinação do conjunto de valores $\phi_i;\ \theta_j;\ i=1,\dots,p; j=1,\dots,q$, que apresente o resultado mais baixo.

4.2.8 Teste de validação em modelos de series temporais

O objectivo na identificação/modelação é retirar de uma determinada série dados gerados por um determinado processo, toda a autocorrelação existente entre os vários instantes de amostragem, até que os respectivos resíduos apresentem uma sequência de ruído branco. Se o ajustamento do modelo é o adequado, a série resultante dos resíduos $\varepsilon_t = Y_t - \hat{Y}_t$ deverá ser ruído branco. Ou seja, o modelo estimado capturou toda a estrutura de autocorrelação da série, peço que não deverá existir qualquer autocorrelação nos resíduos. Assim, ao determinar-se a ACF dos resíduos, esta não deverá ter qualquer componente significante. Uma estatística bastante útil é a estatística Ljung-Box-Pierce, definida por

$$\boldsymbol{Q = n(n+2)\sum_{k=1}^{K}\frac{r_e(k)^2}{n-k}}$$

Esta estatística é utilizada no "teste de Portmanteau", cuja hipótese nula é dada por

$$\boldsymbol{H_0:\ \rho_i = 0,\qquad i = 1,\dots,K}$$

Em que, evidentemente, grande valores de Q são devido a grandes autocorrelações. A hipótese nula é rejeitada se $\boldsymbol{Q > \chi^2_{\alpha,K-p-q}}$.

Utilização dos resíduos para modificar um modelo ARIMA(*p,d,q*)

Se a ACF da série dos resíduos $\{\varepsilon_t\}$ demonstrarem uma estrutura diferente de ruído branco, ou seja, se houver alguma lag que indique alguma autocorrelação, o modelo inicial deverá ser alterado de modo a mover essa autocorrelação. Alternativamente ver Figura 4-6, pode-se ajustar um modelo ARMA(p_e, q_e) aos resíduos (Del Castillo, 2002)

$$\boldsymbol{\widehat{A}_e(\mathcal{B})e_t = \widehat{C}_e(\mathcal{B})\varepsilon_t} \qquad \textbf{(4.110)}$$

O qual, após captura de toda a correlação remanescente deverá produzir uma série $\{\varepsilon_t\}$ contendo apenas ruído branco. O modelo resultante (4.110) deverá então ser combinado com o modelo original ARIMA(*p,d,q*) resultando o modelo

$$\boldsymbol{\widehat{A}_e(\mathcal{B})\widehat{A}(\mathcal{B})\nabla^{\mathrm{d}}\widetilde{Y}_t = \widehat{C}(\mathcal{B})\widehat{C}_e(\mathcal{B})\varepsilon_t} \qquad \textbf{(4.111)}$$

Que é exactamente um processo ARIMA($p+p_e$, d, $q+q_e$).

4.3 Funções de transferência

Na secção 4.1 descreveu-se a função de transferência como sendo uma relação dinâmica entre a entrada e a saída do processo. As variáveis de entrada são as variáveis controláveis, ou pelo menos, parte delas deverão ser controláveis. Como neste ponto apenas se trata de sistemas SISO (Single-input/Single-output), ou seja, sistemas uni-variável, supõe-se que se está perante processos que tem apenas uma variável de entrada, a que se chama variável controlável e uma variável de saída que terá a designação de resposta do processo.

Também se concluiu no ponto 4.1, e que será de extrema importância nesta secção, que a função de transferência de um processo coincide com a resposta do processo a uma entrada em forma de impulso unitário.

4.3.1 Modelos de função de transferência

Nesta secção olha-se para a dinâmica de um sistema como uma caixa preta (Figura 4-7), em que se relaciona a entrada com a saída sem qualquer preocupação com o que se passa no interior da caixa (em contraponto com os modelos de representação em espaço de estados). Nesta secção, as funções de transferência serão essencialmente descritas em termos de equações às diferenças que serão representadas como funções do operador de atraso $\mathcal{B}$.

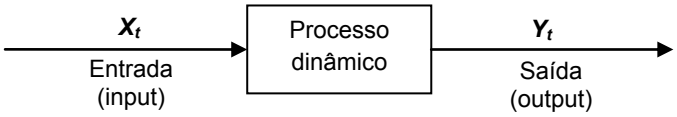


Figura 4-7 Processo dinâmico

O objectivo é determinar relação dinâmica entre as séries $\{X_t\}$ e $\{Y_t\}$, em que os pares $(\{X_t\},\{Y_t\})$ são sequências de observações disponíveis em pontos equidistantes no tempo colectadas de forma sincronizada.

Voltando novamente à resposta ao degrau; assumindo-se que o processo é estável e que a variável de entrada X_t se mantém no nível zero para todo o $t < 0$; se a partir do instante $t = 0$ se fixar o valor da variável de entrada X_t no nível um, então, após um período de transição, a variável de saída Y_t terá atingido a um valor estacionário num nível $Y_\infty = g$. Assumindo-se que o processo é linear, então, para qualquer nível X da variável de entrada, obter-se-á a resposta no estado estacionário dada por

$$\boldsymbol{Y_\infty = gX}$$

Onde g é o ganho estacionário. A ideia por detrás do modelo de resposta ao degrau, é que se pode aplicar uma entrada em degrau unitário à entrada do processo (variável manipulada) e medir a resposta (variável de saída) em anel aberto até esta atingir um valor estacionário. Devido ao facto de se assumir que o processo é linear, o conhecimento da resposta ao degrau unitário permite deduzir a resposta para qualquer outro sinal de entrada. O ganho molda as alterações do estado estacionário da saída que eventualmente seria

obtido devido a uma alteração unitária na entrada. Evidentemente que este é um modelo estático do tipo do modelo obtido através de desenho de experiências com recurso a técnicas de regressão. Por exemplo, quando se implementa um programa de desenho de experiências (DOE) num processo químico, está-se apenas interessado no resultado final, quando o estado estacionário é atingido, descartando-se a fase transiente do processo.

Nos projectos de controlo, o objectivo é fazer com que a saída do processo se mantenha ou regresse rapidamente aos valores desejados/objectivo (*target*), o que implica que a resposta dinâmica evidenciada na fase transiente, seja modelada. Para se modelar a dinâmica do processo, necessita-se de conhecer o comportamento da saída do processo, sempre a que é alterada a entrada do processo. Se o processo for linear, então ter-se-á para resposta do processo uma combinação linear dos valores passados, e possivelmente do valor presente, da entrada do processo:

$$\boldsymbol{Y_t = v_0 X_t + v_1 X_{t-1} + \cdots = \sum_{j=0}^{\infty} v_j X_{t-j} = H(\mathcal{B}) X_t} \qquad \textbf{(4.112)}$$

Onde o polinómio $H(\mathcal{B})$ é de ordem infinita e é definida como a função de transferência do sistema que, como demonstrado na secção 4.1, coincide com a resposta do sistema ao impulso unitário.

Para se ter uma ideia mais intuitiva da função de transferência, considere-se novamente a função impulso unitário (Figura 4-8 direita)

$$\boldsymbol{X_t = \begin{cases} 1 & se\ t = 0 \\ 0 & se\ t \neq 0 \end{cases}}$$

A resposta do sistema ao impulso unitário g_t, será dada pelos pesos $v_0, v_1, \ldots$ (Figura 4-8 esquerda), que coincide com a função de transferência porque se a entrada, num determinado instante, corresponder a um impulso de magnitude X, a resposta será a resposta ao impulso unitário multiplicada pela magnitune do impulso de entrada X.

Figura 4-8 Função impulso (a) e resposta ao impulso g_t (b)

Considerando-se um sinal de entrada como uma sequência de impulso de magnitude X_t, a resposta a esse sinal será dado pelo somatório de convulsão das respostas individuais de cada impulso, ou pelo integral de convulsão (ver secção 4.1.1.3) no caso da análise continua. Suponha-se que ao sistema descrito pela função de transferência da Figura 4-8 que lhe é impostos um nível de entrada nulo para $t < 0$, e que nos três instantes seguintes, o sinal de entrada "discretizado" (em forma de impulso) tem a forma representada na Figura 4-9 (lado esquerdo – input). No lado direito da Figura 4-9 estão representadas as respostas a cada um dos impulsos sequenciais juntamente com a resposta global do sistema. É fácil

verificar que a resposta global do sistema (Figura 4-9 a preto) será dada pela soma acumulada de cada uma das respostas para o mesmo instante, ou seja, por exemplo para $t = 2$, tem-se $Y_2 = \sum_{j=0}^{\infty} v_j \, X_{t-j} = v_0 X_2 + v_1 X_1 + v_2 X_0$.

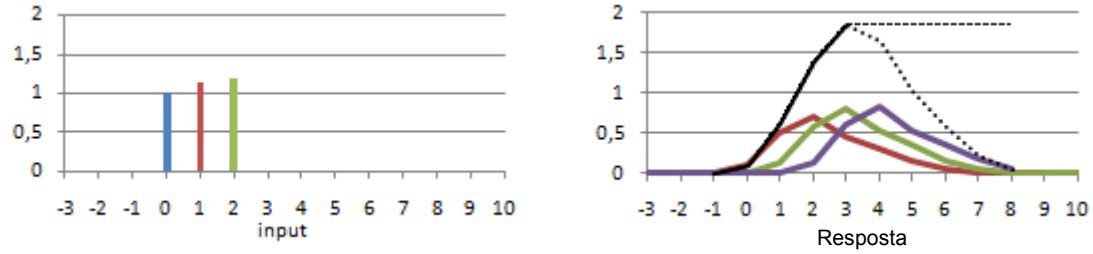


Figura 4-9 Sinal de entrada (esquerda) e respectiva resposta

Forma incremental

Definindo-se

$$\boldsymbol{y_t = Y_t - Y_{t-1} = (1 - \mathcal{B})Y_t = \nabla Y_t}$$

e

$$\boldsymbol{x_t = X_t - X_{t-1} = (1 - \mathcal{B})X_t = \nabla X_t}$$

os incrementos em X e Y. Diferenciando (4.112) obtém-se

$$\boldsymbol{y_t = H(\mathcal{B})x_t}$$

Ou seja, a forma incremental satisfaz o mesmo modelo de função de transferência.

Estabilidade

Se uma série infinita $v_0 + v_{01}\mathcal{B} + v_1\mathcal{B}^2 + \cdots$ converge para $|\mathcal{B}| \leq 1$, ou equivalentemente, se os v_j são absolutamente somáveis, tal que $\sum_{j=0}^{\infty} |v_j| < \infty$ então o diz-se que o sistema é estável. A estabilidade implica que um incremento finito na entrada resulta num incremento finito na saída.

A condição de estabilidade significa que os pesos v_j deverão tender para zero, ou serem iguais a zero, para grandes valores de t, o que implica que o efeito de entradas passadas em respostas futuras é desprezível ou eventualmente nulo a partir de determinado tempo.

Função de transferência na forma de divisão de polinómios

A função de transferência escrita na forma (4.112) é composta por uma série infinita, o que não é prático. Há toda a conveniência em escrever a função de transferência numa forma fechada, ou seja, uma representação mais "parcimónia" representada pela forma geral de equação às diferenças

$$\boldsymbol{Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \cdots + a_r Y_{t-r} + b_0 X_{t-k} - b_1 X_{t-k-1} - \cdots - b_s X_{t-k-s}}$$

ou

$$\boldsymbol{A_r(\mathcal{B})Y_t = B_s(\mathcal{B})X_{t-k} = B_s(\mathcal{B})\mathcal{B}^k X_t}$$

Onde $A_r(\mathcal{B})$ é um polinómio em $\mathcal{B}$ de ordem r e $B_s(\mathcal{B})$ é um polinómio de ordem s. Ao desfasamento k dá-se o nome de atraso entrada resposta (*input-output delay*) e corresponde ao tempo que o sistema demora a responder a uma determinada entrada.

Comparando as duas representações da função de transferência, chega-se à chamada representação (*r,s,k*) da função de transferência

$$\boldsymbol{H(\mathcal{B}) = \frac{B_s(\mathcal{B})\mathcal{B}^k}{A_r(\mathcal{B})}} \quad \textbf{(4.113)}$$

O modelo ARIMA $A(\mathcal{B})z_t = C(\mathcal{B})\varepsilon_t$ usado para representar séries temporais $\{z_t\}$ relaciona $z_t \text{com } \varepsilon_t$ através do filtro linear

$$\boldsymbol{z_t = A^{-1}(\mathcal{B})C(\mathcal{B})\varepsilon_t}$$

onde ε_t é ruído branco. Verifica-se assim que um modelo ARIMA pode ser utilizado para representar a saída de um sistema dinâmico cuja entrada é ruído branco e em que a função de transferência pode parcimoniosamente expressa como razão de dois operadores polinomiais em $\mathcal{B}$.

Devido ao paralelismo entre a função de transferência discreta e o modelo ARMA, as condições de estabilidade para a função de transferência discreta coincidem com as condições impostas ao modelo ARMA, ou seja, que as raízes do polinómio $A_r(\mathcal{B}) = 0$ se situem estritamente fora do circulo unitário.

Para um sistema estável, o ganho estacionário pode ser determinado a partir da função de transferência na forma racional aplicando uma entrada $X_t = 1$ para todo o t, obtendo-se

$$\boldsymbol{\lim_{t\to\infty} Y_t = Y_\infty = \frac{b_0 - b_1 - \cdots - b_s}{1 - a_1 - a_2 - \cdots - a_r} = g}$$

Este resultado está directamente relacionado com o teorema do Valor final (Teorema do valor final - ver secção 4.1.2.5) da transformada $\mathcal{Z}$ (Del Castillo, 2002).

4.3.2 Identificação de funções de transferência de processos

Na identificação da função de transferência de um processo, a primeira etapa consiste em determinar valores referentes às ordens r e s e ao atraso entrada-saída k do modelo racional, que determinam a sua estrutura. Uma das ferramentas proposta por (Box, Jenkins, & Reinsel, 2008), entre outros, para alcançar este objectivo, é a resposta ao degrau.

Se as duas representações da função de transferência, resposta ao impulso (4.112), e racional (4.113), então elas são equivalentes, ou seja, verifica-se a igualdade

$$\boldsymbol{v_0 + v_1\mathcal{B} + v_2\mathcal{B}^2 + \cdots = \frac{(b_0 - b_1\mathcal{B} - \cdots - b_s\mathcal{B}^s)\mathcal{B}^k}{1 - a_1\mathcal{B} - \cdots - a_r\mathcal{B}^r}}$$

Que pode ser reescrita na forma

$$(\mathbf{1} - \boldsymbol{a_1}\mathcal{B} - \cdots - \boldsymbol{a_r}\mathcal{B}^r)\left(\boldsymbol{v_0} + \boldsymbol{v_1}\mathcal{B} + \boldsymbol{v_2}\mathcal{B}^2 + \cdots\right) = (\boldsymbol{b_0} - \boldsymbol{b_1}\mathcal{B} - \cdots - \boldsymbol{b_s}\mathcal{B}^s)\mathcal{B}^k \quad \textbf{(4.114)}$$

A análise desta relação poderá ser bastante util para a identificação de r, s e k. Igualando os coeficientes de $\mathcal{B}$, chega-se à relação (Box, Jenkins, & Reinsel, 2008)

$$\boldsymbol{v_j} = \begin{cases} \mathbf{0} & \boldsymbol{j} < k \\ \boldsymbol{a_1 v_{j-1}} + \boldsymbol{a_2 v_{j-2}} + \cdots + \boldsymbol{a_r v_{j-r}} + \boldsymbol{b_0} & \boldsymbol{j} = \boldsymbol{k} \\ \boldsymbol{a_1 v_{j-1}} + \boldsymbol{a_2 v_{j-2}} + \cdots + \boldsymbol{a_r v_{j-r}} - \boldsymbol{b_{j-b}} & \boldsymbol{j} = \boldsymbol{k+1}, \boldsymbol{k+2}, \ldots, \boldsymbol{k+s} \\ \boldsymbol{a_1 v_{j-1}} + \boldsymbol{a_2 v_{j-2}} + \cdots + \boldsymbol{a_r v_{j-r}} & \boldsymbol{j} > k+s \end{cases} \quad \textbf{(4.115)}$$

Os pesos $v_{k+s}, v_{k+s-1}, \ldots, v_{k+s-r-1}$ fornecem os valores de partida da equação às diferenças

$$\boldsymbol{A(\mathcal{B})v_j} = \mathbf{0} \qquad \boldsymbol{j} > k+s$$

A solução $v_j = f(a, b, j)$ desta equação às diferenças aplica-se a todos os valores v_j para os quais $j \geq k + s - r + 1$.

Em geral, os pesos v_j da resposta ao impulso consistem em

1. k valores iguais a zero, $v_0, v_1, \ldots, v_{k-1}$
2. $s - r + 1$ valores adicionais, $v_k, v_{k+1}, \ldots, v_{k+s-r}$, que não seguem um padrão fixo (se $s < r$ estes valores não existem)
3. Valores v_j com $j \geq k + s - r + 1$ seguindo um padrão de acordo com a ordem r da equação às diferenças, os quais incluem os r valores de partida $v_{k+s}, v_{k+s-1}, \ldots, v_{k+s-r-1}$. Os valores de partida v_j para $j < k$ serão nulos.

Resposta ao degrau

A função resposta ao impulso $v(\mathcal{B})$ pode obter-se através da diferenciação da corresponde resposta ao degrau $V(\mathcal{B})$

$$\boldsymbol{v(\mathcal{B})} = (\mathbf{1} - \boldsymbol{\mathcal{B}})\boldsymbol{V(\mathcal{B})} \quad \textbf{(4.116)}$$

Em que

$$\begin{aligned} \boldsymbol{V(\mathcal{B})} &= \boldsymbol{V_0} + \boldsymbol{V_1}\mathcal{B} + \boldsymbol{V_2}\mathcal{B}^2 + \cdots \\ &= \boldsymbol{v_0} + (\boldsymbol{v_0} + \boldsymbol{v_1})\mathcal{B} + +(\boldsymbol{v_0} + \boldsymbol{v_1} + \boldsymbol{v_2})\mathcal{B}^2 + \cdots \end{aligned} \quad \textbf{(4.117)}$$

Substituindo (4.116) em (4.114) obtém-se a identidade

$$\begin{aligned} &\left(\mathbf{1} - \boldsymbol{a^*}_1\mathcal{B} - \cdots - \boldsymbol{a^*}_r\mathcal{B}^{r+1}\right)\left(\boldsymbol{v_0} + \boldsymbol{v_1}\mathcal{B} + \boldsymbol{v_2}\mathcal{B}^2 + \cdots\right) \\ &\qquad = (\boldsymbol{b_0} - \boldsymbol{b_1}\mathcal{B} - \cdots - \boldsymbol{b_s}\mathcal{B}^s)\mathcal{B}^k \end{aligned} \quad \textbf{(4.118)}$$

com

$$\left(\mathbf{1} - \boldsymbol{a^*}_{\mathbf{1}}\boldsymbol{\mathcal{B}} - \cdots - \boldsymbol{a^*}_{\boldsymbol{r}}\boldsymbol{\mathcal{B}}^{\boldsymbol{r+1}}\right) = (\mathbf{1} - \boldsymbol{\mathcal{B}})(\mathbf{1} - \boldsymbol{a_1}\boldsymbol{\mathcal{B}} - \cdots - \boldsymbol{a_r}\boldsymbol{\mathcal{B}}^{\boldsymbol{r}}) \quad \textbf{(4.119)}$$

Verifica-se que identidade (4.118) para os pesos $\boldsymbol{V_j}$ da resposta ao degrau é "paralela" à identidade (4.114) para os pesos da resposta ao impulso, excepto que o operador do lado direito $\boldsymbol{A^*(\mathcal{B})}$ é de ordem $r+1$ em vez de r (Box, Jenkins, & Reinsel, 2008)

Utilizando o resultado (4.115), conclui-se que a função resposta ao degrau é definida por:

1. k valores iguais a zero, $V_0, V_1, \ldots, V_{k-1}$
2. $s-r$ valores adicionais, $V_k, V_{k+1}, \ldots, V_{k+s-r-1}$, que não seguem um padrão fixo (se $s < r$ estes valores não existem)
3. Valores V_j com $j \geq k+s-r$ seguindo um padrão de acordo com a ordem $r+1$ da equação às diferenças $A^*(\mathcal{B})V_j = 0$, os quais incluem os $r+1$ valores de partida $V_{k+s}, V_{k+s-1}, \ldots, V_{k+s-r}$. Os valores de partida V_j para $j < k$ serão nulos.

Conhecendo-se o comportamento típico das respostas ao impulso $\{v_j\}$ e/ou respostas ao degrau $\{V_j\}$, é possível identificar r, s e k de um determinado sistema a partir das respectivas observações. (Del Castillo, 2002, pp. 140-141) e (Box, Jenkins, & Reinsel, 2008, pp. 451-453) entre outros, contém tabelas que ilustram os comportamentos típicos das funções de resposta ao impulso e resposta ao degrau para diferentes valores de r, s e k.

4.3.2.1 Identificação de funções de transferência na presença de ruído

Na secção anterior introduziu-se o estudo de funções de transferência discretas do tipo

$$\boldsymbol{Y_t} = \frac{\boldsymbol{B(\mathcal{B})X_{t-k}}}{\boldsymbol{A(\mathcal{B})}}$$

Neste modelo X_t e Y_t são desvios relativamente ao equilíbrio das entradas e saídas dos sistemas. Na prática, os sistemas estão sempre infectados por distúrbios ou ruído cujos efeitos práticos se traduzem na corrupção dos valores esperados para das saídas dos sistemas por uma quantidade N_t, pelo que a identificação em tais condições apenas é possível se o nível de ruído é moderado em comparação com o nível do sinal que se pretende identificar.

Quando o nível dos distúrbios N_t é significativo comparativamente ao sinal, este normalmente entra no processo como um parâmetro não controlável, em contraste com X_t. Para modelar os distúrbios, (Box, Jenkins, & Reinsel, 2008) propõe a modelação de N_t por um processo ARIMA(*p,d,q*) e a sua inclusão no modelo anterior

$$\boldsymbol{Y_t} = \frac{\boldsymbol{B(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}}\boldsymbol{X_{t-k}} + \boldsymbol{N_t}$$

Onde

$$\boldsymbol{N_t} = \frac{\boldsymbol{C(\mathcal{B})}}{\boldsymbol{D(\mathcal{B})}}\boldsymbol{\varepsilon_t}$$

em que $D(\mathcal{B})$ pode ter uma ou mais raízes unitárias. Este modelo é conhecido na literatura de engenharia de controlo como modelo Box-Jenkins, ou modelo BJ (Figura 4-10):

$$\boldsymbol{Y_t = \frac{B(\mathcal{B})}{A(\mathcal{B})} X_{t-k} + \frac{C(\mathcal{B})}{D(\mathcal{B})} \varepsilon_t} \quad \textbf{(4.120)}$$

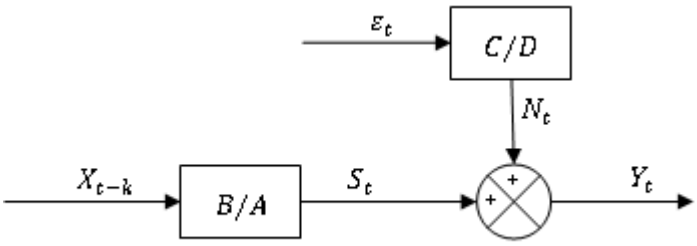


Figura 4-10 Função de transferência-ruido Box-Jenkins

As ferramentas utilizadas em engenharia de controlo para identificação das funções de transferência são normalmente baseadas em determinadas escolhas de entradas do sistema, como por exemplo impulsos, degraus ou ondas sinusoidais. Estas metodologias são extremamente úteis quando os distúrbios são relativamente reduzidos em relação aos sinais de entrada e saída. Quando os distúrbios são significantes torna-se necessário o uso de técnicas estatísticas para estimar a função de transferência. Para efeitos de modelação, (Box, Jenkins, & Reinsel, 2008) propõe que a entrada do sistema X_t seja assumida como um processo estocástico e não como uma variável controlada. Pretende-se com isto dizer que, quando o sistema está a correr com objectivos de identificação/modelação, a variável de entrada deve comportar-se como uma variável aleatória percorrendo o máximo de valores admissíveis de modo a avaliar a resposta do sistema nos mais diversos níveis e transições, pelo que não deve existir qualquer acção de controlo em que a variável de entrada seja manipulada.

Abordagem a sistemas multivariados

Do mesmo modo que a função de autocorrelação é utilizada para identificar modelos estocásticos, uma das ferramentas utilizadas para modelação de funções de transferência de processos é a função de correlação cruzada entre a entrada e saída. Suponha-se que se pretende descrever uma série temporal X_t da entrada de um sistema físico, e a correspondente resposta Y_t. Então este par de séries temporais pode ser visto como uma única série temporal em que cada observação é constituída por vector bidimensional à qual se pode chamar "*processo estocástico bivariado*" (X_t, Y_t). Deve-se assumir que os dados são lidos simultaneamente e em intervalos equidistantes resultando um par de valores de uma série temporal discreta, gerado por um processo bivariado, em que os valores da série nos instantes $t_0 + h, t_0 + 2h, \dots, t_0 + Nh$ são descritos por $(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)$.

Funções de covariância cruzada e correlação cruzada

Em geral, um processo estocástico bivariado (X_t, Y_t) não necessita de ser estacionário. Contudo, como anteriormente, assume-se que um apropriado grau de diferenciação conduz a um processo (x_t, y_t), onde $x_t = \nabla^{d_x} X_t$ e $y_t = \nabla^{d_y} Y_t$ são estacionários. Como anteriormente, o conceito de estacionaridade implica que os processos constituintes x_t e y_t tem média constante μ_x e μ_y e variâncias constantes σ_x^2 e σ_y^2. Se adicionalmente assumir-se que o processo bivariado é normal, ou Gaussiano, ele será caracterizado unicamente pelo vector das suas médias $(\mu_x\ ,\ \mu_y)$ e a sua matriz das covariâncias.

Os coeficientes de covariância de cada uma das séries constituintes na *lag k* são definidos pela fórmula usual:

$$\boldsymbol{\gamma_{xx}(k) = E[(x_t - \mu_x)(x_{t+k} - \mu_x)] = E[(x_t - \mu_x)(x_{t-k} - \mu_x)]}$$

$$\boldsymbol{\gamma_{yy}(k) = E\left[(y_t - \mu_y)(y_{t+k} - \mu_y)\right] = E\left[(y_t - \mu_y)(y_{t-k} - \mu_y)\right]}$$

Onde se utiliza a notação $\boldsymbol{\gamma_{xx}(k)}$ e $\boldsymbol{\gamma_{yy}(k)}$ para a autocovariância das séries x_t e y_t que constituem a diagonal da matriz das covariâncias. Os restantes coeficientes da matriz de covariância são as covariâncias cruzadas entre as séries x_t e y_t na *lag +k:*

$$\boldsymbol{\gamma_{xy}(k) = E\left[(x_t - \mu_x)(y_{t+k} - \mu_y)\right]} \qquad \boldsymbol{k = 0, 1, 2, \ldots} \tag{4.121}$$

e as covariâncias cruzadas entre as séries y_t e x_t na lag +k:

$$\boldsymbol{\gamma_{yx}(k) = E\left[(y_t - \mu_y)(x_{t+k} - \mu_x)\right]} \qquad \boldsymbol{k = 0, 1, 2, \ldots} \tag{4.122}$$

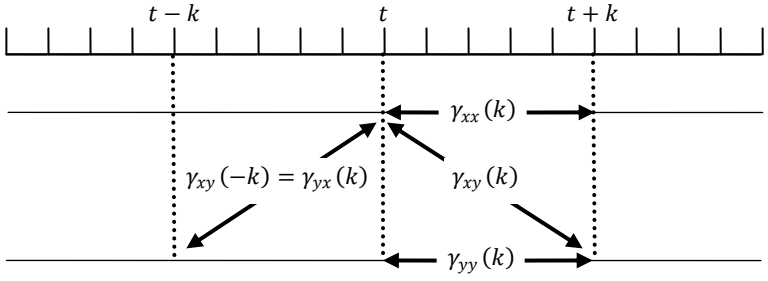


Figura 4-11 - Autocovariância e covariância cruzada de um processo estocástico bivariado

De forma similar ao caso univariado, em condições de estacionariridade (bivariada), a covariância cruzada deve ser a mesma para todo o t, pelo que deve ser função apenas da lag k.

Generalizando-se para série multivariadas, pode dizer-se que a matriz de covariância é composta por

$$\boldsymbol{\gamma_{ij}(k) = E\left[(x_i - \mu_i)(x_{j,t+k} - \mu_j)\right]} \qquad \boldsymbol{k = 0, 1, 2, \ldots}$$

Em termos de covariância cruzada, verifica-se que $\gamma_{yx}(k)$ não é igual a $\gamma_{xy}(k)$. Mas como (ver Figura 4-11):

$$\boldsymbol{\gamma_{xy}(k) = E\left[(x_{t-k} - \mu_x)(y_t - \mu_y)\right] = E\left[(y_t - \mu_{xy})(x_{t-k} - \mu_x)\right] = \gamma_{yx}(-k)}$$

Apenas se necessita de definir uma função $\boldsymbol{\gamma_{xy}(k)}$ para $k = 0, \pm 1, \pm 2, \ldots$. Tal como já foi dito acima, a função $\boldsymbol{\gamma_{xy}(k) = cov[x_t, y_{t+k}]}$, como definida em (4.121) chama-se função de correlação cruzada de um processo bivariado estacionário, a qual pode também ser representada de forma matricial, como se verá mais adiante

$$\boldsymbol{\Gamma(k) = \begin{bmatrix} \gamma_{xx}(k) & \gamma_{yx}(k) \\ \gamma_{xy}(k) & \gamma_{yy}(k) \end{bmatrix}}$$

De forma similar, à correlação entre x_t e y_{t+k} que é dada pela quantidade adimensional definida por

$$\boldsymbol{\rho_{xy}(k) = \frac{\gamma_{xy}(k)}{\sigma_x \sigma_y} \qquad k = 0, 1, 2, \ldots} \tag{4.123}$$

dá-se o nome de coeficiente de correlação cruzada da lag k, sendo atribuído o nome de função de correlação cruzada de um processo bivariado estacionário à função $\rho_{xy}(k)$, definida para $k = 0, \pm 1, \pm 2, \ldots$ (Box, Jenkins, & Reinsel, 2008).

Em contraste com a função de autocorrelação, geralmente a função de autocorrelação cruzada não é simétrica.

Estimação das funções de autocorrelação e autocovariância cruzadas

Após a diferenciação *d* vezes das séries originais de entrada e saída, até que estas mostrem comportamento estacionário, obter-se-á $n = N - d$ pares de valores (x_1, y_1), (x_2, y_2), …, (x_n, y_n) disponíveis para análise. A partir desses valores será então possível obter uma estimativa $c_{xy}(k)$ dos coeficientes da covariância cruzada da *lag k* dada por

$$\boldsymbol{c_{xy}(k) = \hat{\gamma}_{xy}(k) = \begin{cases} \frac{1}{n}\sum_{t=1}^{n-k}(x_t - \bar{x})(y_{t+k} - \bar{y}) & k = 0, 1, 2, \ldots \\ \frac{1}{n}\sum_{t=1}^{n+k}(y_t - \bar{y})(x_{t-k} - \bar{x}) & k = 0, -1, -2, \ldots \end{cases}} \tag{4.124}$$

Onde $\bar{x}$ e $\bar{y}$ são as médias amostrais de das séries x_t e y_t respectivamente. De forma similar, as estimativas $r_{xy}(k)$ dos coeficientes de correlação cruzada $\rho_{xy}(k)$ para a lag k podem obter-se através da substituição em (4.123) de $\gamma_{xy}(k)$ pelas respectivas estimativas $c_{xy}(k)$, σ_x por $s_x = \sqrt{c_{xx}(0)}$ e σ_y por $s_y = \sqrt{c_{yy}(0)}$:

$$\boldsymbol{r_{xy}(k) = \frac{c_{xy}(k)}{s_x s_y} \qquad k = 0, \pm 1, \pm 2, \ldots} \tag{4.125}$$

Um teste primitivo para verificar se certos valores da função de correlação cruzada $\boldsymbol{\rho_{xy}(k)}$, serão efectivamente zero pode ser efectuado através da comparação das correspondentes estimativas com os respectivos desvios padrão a partir da formula de Bartlett.

Bartlett mostrou que, se as séries x_t e y_t não estão correlacionadas, então (Del Castillo, 2002):

$$\boldsymbol{var\left[r_{xy}(k)\right] \approx \frac{1}{n}\sum_{-\infty}^{\infty}\rho_x(k)\rho_y(k) \qquad k = 0, 1, 2, \ldots}$$

Daí que, se duas séries são não correlacionadas e uma é ruído branco ($\rho(k) = 0$, qualquer que seja k), então verifica-se que

$$\boldsymbol{var}\left[\boldsymbol{r_{xy}(k)}\right] = \boldsymbol{var}\left[\boldsymbol{r_{xy}(-k)}\right] \approx \frac{\boldsymbol{1}}{\boldsymbol{n-k}}$$

Assim, se as séries x_t e x_t são não correlacionadas e uma é ruído branco, a correlação cruzada varia em torno de zero com desvio padrão aproximadamente igual a $(n-k)^{-1/2}$.

Identificação de modelos de funções de transferência

Suponha-se que o modelo da função de transferência

$$\boldsymbol{Y_t = v(\mathcal{B})X_t + N_t} \quad \textbf{(4.126)}$$

Pode ser parametrizado de forma parcimoniosa na forma

$$\boldsymbol{Y_t = A^{-1}(\mathcal{B})B(\mathcal{B})X_{t-k} + N_t} \quad \textbf{(4.127)}$$

O procedimento de identificação é composto pelas seguintes etapas:

1. Determinar estimativas preliminares dos pesos $\hat{v}_j$ da função resposta ao impulso no para o formato (4.126).
2. Utilizar as estimativas obtidas no ponto anterior para tentar estimar as ordens r e s dos operadores do denominador e numerador na equação (4.127) e o parâmetro de atraso k.
3. Substituir os valores $\hat{v}_j$ estimados na equação (4.115), de acordo com os valores r, s e b obtidos no ponto anterior para estimar os parâmetros A e B da equação (4.127).

Uma vez estimados os valores $\hat{v}_j$, é possível ter uma ideia das ordens r e s dos operadores do denominador e numerador na equação (4.127) e do parâmetro de atraso k comparando o andamento de $\hat{v}_j$, com as diversas respostas ao impulso tabeladas em (Del Castillo, 2002, pp. 140-141) e (Box, Jenkins, & Reinsel, 2008, pp. 451-453). Uma vez estimados os valores de r, s e k é então possível determinar os coeficientes a_i b_j através das equações (4.115).

Identificação de funções de transferência pelo branqueamento da entrada

Pelo que foi dito acima, torna-se claro que a correlação dentro de cada uma das séries pode esconder a correlação entre elas, pelo que a identificação dos processos pode ser facilitada se a entrada do sistema na fase de identificação for constituído unicamente por ruído branco. De facto, sempre que se pode escolher a entrada do sistema a identificar, é de todo recomendável que se opte por ruído branco. Quando tal não é possível, e a entrada seguir outro processo estocástico, uma hipótese a ter em conta será recorrer ao branqueamento da série de entrada, conseguindo-se assim uma série composta por ruído branco.

Suponha-se que um processo de entrada x_t, adequadamente diferenciado, é estacionário e pode ser adequadamente representado por um modelo ARMA(p,q) genérico. Então, a partir de um conjunto de dados, é possível construir o modelo do processo x_t

$$\boldsymbol{A_x(\mathcal{B})x_t = C_x(\mathcal{B})\alpha_t}$$

onde α_t é ruído branco. Resolvendo em ordem a α_t obtém-se o filtro que irá permitir branquear a série x_t

$$\boldsymbol{\alpha_t = C_x^{-1}(\mathcal{B})A_x(\mathcal{B})x_t}$$

Aplicando o mesmo filtro à serie de saída y_t obtém-se

$$\boldsymbol{\beta_t = C_x^{-1}(\mathcal{B})A_x(\mathcal{B})y_t}$$

Onde β_t não é necessariamente ruído branco. Substituindo em (4.120), a função de transferência-ruído BJ pode ser reescrita na forma

$$\underbrace{\frac{\boldsymbol{A_x(\mathcal{B})}}{\boldsymbol{C_x(\mathcal{B})}}\boldsymbol{Y_t}}_{\boldsymbol{\beta_t}} = \underbrace{\frac{\boldsymbol{B(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}}\boldsymbol{\mathcal{B}^k}}_{\boldsymbol{H(\mathcal{B})}}\underbrace{\frac{\boldsymbol{A_x(\mathcal{B})}}{\boldsymbol{C_x(\mathcal{B})}}\boldsymbol{X_{t-k}}}_{\boldsymbol{\alpha_t}} + \underbrace{\frac{\boldsymbol{A_x(\mathcal{B})}}{\boldsymbol{C_x(\mathcal{B})}}\boldsymbol{N_t}}_{\boldsymbol{\varepsilon_t^*}}$$

ou

$$\boldsymbol{\beta_t = H(\mathcal{B})\alpha_t + \varepsilon_t^*}$$

Onde $\varepsilon_t^* = C_x^{-1}(\mathcal{B})A_x(\mathcal{B})N_t$ é ruído colorido. Multiplicando-se ambos os lados por α_{t-k} e tomando-se o valor esperado obtém-se

$$\boldsymbol{E[\alpha_{t-k}\,\beta_t] = E[\alpha_{t-k}\,H(\mathcal{B})\alpha_t] + E[\alpha_{t-k}\,\varepsilon_t^*]}$$

ou seja

$$\boldsymbol{\gamma_{\alpha\beta}(k) = E\left[\alpha_{t-k}v_k\mathcal{B}^k\alpha_t\right] + E[\alpha_{t-k}\,\varepsilon_t^*]}$$

Onde $\gamma_{\alpha\beta}(\boldsymbol{k}) = E[\alpha_{t-k}\,\beta_t]$ é a covariância cruzada para a lag k entre as séries α_t e β_t. Desta equação tira-se que:

1. Os valores da série $\boldsymbol{\beta_t}$ estão relacionados com valores passados $\boldsymbol{t-k}$ da série $\boldsymbol{\alpha_t}$, ou seja, a saída do processo depende dos valores passados, e talvez presente, da entrada, pelo que o membro direito será diferente de zero;
2. No primeiro termo do membro direito, como $\boldsymbol{\alpha_t}$ é ruído branco, tem-se que $E[\alpha_{t-k}v_k\mathcal{B}^k\alpha_t] = v_kE[\alpha_{t-k}\mathcal{B}^k\alpha_t] = v_k\sigma_\alpha^2$;
3. O segundo termo do membro direito é zero, porque $\boldsymbol{\varepsilon_t^*}$ não é função de $\boldsymbol{\alpha_t}$, e $\boldsymbol{\alpha_t}$ e $\boldsymbol{\varepsilon_t^*}$ não estão correlacionados.

Das equações anteriores resulta então que

$$\boldsymbol{\gamma_{\alpha\beta}(k) = v_k\sigma_\alpha^2} \quad \Rightarrow \quad \boldsymbol{v_k = \frac{\gamma_{\alpha\beta}(k)}{\sigma_\alpha^2}}$$

Ou em termos de correlação cruzada

$$\boldsymbol{v_k = \frac{\rho_{\alpha\beta}(k)\sigma_\beta}{\sigma_\alpha}} \qquad \boldsymbol{k = 0, 1, 2, \ldots} \tag{4.128}$$

Verifica-se assim que, após o branqueamento da entrada, a função de correlação cruzada entre as duas séries transformadas é directamente proporcional à função resposta ao impulso.

Na prática, não é necessário conhecer a função de correlação cruzada teórica $\boldsymbol{\rho_{\alpha\beta}(k)}$, porque a estimativa da função resposta ao impulso é dada directamente através dos valores estimados das variáveis envolvidas na equação anterior, ou seja

$$\boldsymbol{\hat{v}_k = \frac{r_{\alpha\beta}(k)s_\beta}{s_\alpha}} \qquad \boldsymbol{k = 0, 1, 2, \ldots} \tag{4.129}$$

Esta é a função resposta ao impulso estimada, a partir da qual se podem identificar os valores referentes a (*r,s,k).*

Nota 1: Se existir qualquer tipo de controlo por feedback, o que vai acontecer é que os valores de entrada x_t vão ser determinados a partir dos valores da saída y_t, pelo que existirá sempre alguma correlação entre a série de entrada x_t a série de saída y_t. Por outro lado, se a série de entrada estiver correlacionada com a série de saída, torna-se difícil determinar se a correlação dentro da série de saída se deve à própria dinâmica do processo ou ao facto de a série de entrada estar correlacionada com a série de saída.

Nota 2: Há no entanto um problema que se levanta com o branqueamento. Nos sistemas físicos reais, os valores das séries de entrada são referentes aos valores retirados dos actuadores, que normalmente têm também alguma dinâmica. O que se pretende normalmente é que a constante de tempo referente aos actuadores seja insignificante comparativamente com a constante de tempo da dinâmica do processo. Ao proceder-se ao branqueamento da série de entrada, está-se simultaneamente a branquear a dinâmica dos actuadores

Identificação do modelo do ruído

Voltando ao caso geral, suponha-se que o modelo pode ser escrito na forma

$$\boldsymbol{y_t = v(\mathcal{B})x_t + n_t}$$

Onde $n_t = \nabla^{\mathrm{d}} N_t$. Após a obtenção da função de transferência, a serie dos resíduos pode obter-se a partir da equação

$$\boldsymbol{\hat{n}_t = y_t - \hat{v}(\mathcal{B})x_t}$$

Alternativamente, o operador $\hat{v}(\mathcal{B})$ pode ser substituído pelo modelo de função de transferência estimado $\hat{A}^{-1}(\mathcal{B})\hat{B}(\mathcal{B})\mathcal{B}^k$ determinado a partir da identificação preliminar

$$\widehat{\boldsymbol{n}}_{\boldsymbol{t}} = \boldsymbol{y}_{\boldsymbol{t}} - \widehat{\boldsymbol{A}}^{-1}(\boldsymbol{\mathcal{B}})\widehat{\boldsymbol{B}}(\boldsymbol{\mathcal{B}})\boldsymbol{x}_{\boldsymbol{t-k}}$$

A série temporal dos resíduos $\widehat{\boldsymbol{n}}_{\boldsymbol{t}}$ obtida poderá então ser ajustada a um modelo ARIMA(p,*d*,*q*) com recurso a técnicas com as funções de autocorrelação e autocorrelação parcial discutidas em pontos anteriores.

Modelo ARMAX

O modelo de função de transferência Box-Jenkins considera dinâmicas distintas para o modelo do sistema e das perturbações. Suponha-se que um particular sistema é representado pelo modelo

$$\boldsymbol{y}_{\boldsymbol{t}} = \frac{\boldsymbol{B}^*(\boldsymbol{\mathcal{B}})}{\boldsymbol{A}^*(\boldsymbol{\mathcal{B}})}\boldsymbol{x}_{\boldsymbol{t-k}} + \frac{\boldsymbol{C}^*(\boldsymbol{\mathcal{B}})}{\boldsymbol{D}^*(\boldsymbol{\mathcal{B}})}\boldsymbol{\varepsilon}_{\boldsymbol{t}}$$

Então este poderá ser igualmente representado na forma

$$\boldsymbol{A}(\boldsymbol{\mathcal{B}})\boldsymbol{y}_{\boldsymbol{t}} = \boldsymbol{A}(\boldsymbol{\mathcal{B}})\frac{\boldsymbol{B}^*(\boldsymbol{\mathcal{B}})}{\boldsymbol{A}^*(\boldsymbol{\mathcal{B}})}\boldsymbol{x}_{\boldsymbol{t-k}} + \boldsymbol{A}(\boldsymbol{\mathcal{B}})\frac{\boldsymbol{C}^*(\boldsymbol{\mathcal{B}})}{\boldsymbol{D}^*(\boldsymbol{\mathcal{B}})}\boldsymbol{\varepsilon}_{\boldsymbol{t}}$$

onde $A(\mathcal{B})$ pode ser um qualquer factor comum, e portanto será redundante. Se o factor $\boldsymbol{A}(\mathcal{B})$ for igual a $A^*(\mathcal{B})D^*(\mathcal{B})$ então obter-se-á

$$\boldsymbol{A}^*(\boldsymbol{\mathcal{B}})\boldsymbol{D}^*(\boldsymbol{\mathcal{B}})\boldsymbol{y}_{\boldsymbol{t}} = \boldsymbol{D}^*(\boldsymbol{\mathcal{B}})\boldsymbol{B}^*(\boldsymbol{\mathcal{B}})\boldsymbol{x}_{\boldsymbol{t-k}} + \boldsymbol{A}^*(\boldsymbol{\mathcal{B}})\boldsymbol{C}^*(\boldsymbol{\mathcal{B}})\boldsymbol{\varepsilon}_{\boldsymbol{t}}$$

Que também poderá ser escrito na forma

$$\boldsymbol{A}(\boldsymbol{\mathcal{B}})\boldsymbol{y}_{\boldsymbol{t}} = \boldsymbol{B}(\boldsymbol{\mathcal{B}})\boldsymbol{x}_{\boldsymbol{t-k}} + \boldsymbol{C}(\boldsymbol{\mathcal{B}})\boldsymbol{\varepsilon}_{\boldsymbol{t}} \quad \textbf{(4.130)}$$

Este modelo é referido como modelo ARMAX, cujo nome vem da estrutura ARMA mais a variável exógena. Neste modelo, a variável exógena é X_t e a variável endógena é Y_t. Esta forma torna-se mais simples de manipular quando os objectivos do modelo estão relacionados com projecto de controladores como se verá em capítulos posteriores.

4.3.2.2 Ajustamento e validação de modelos de funções de transferência

Até aqui tratou-se de fazer uma estimativa grosseira da função de transferência-ruido Box-Jernkins a um conjunto sequencial de pares de observações (X_t, Y_t) retiradas de um determinado processo de forma síncrona e em instantes equidistantes (Deve-se utilizar no mínimo 100 pares (Del Castillo, 2002)). O próximo passo será tentar estimar simultaneamente e eficientemente os parâmetros k, $A = (a_1, a_2, \dots, a_r)$, $B = (b_1, b_2, \dots, b_s)$, $C = (c_1, c_2, \dots, c_p)$ e $D = (a_1, a_2, \dots, a_q)$ do modelo

$$\boldsymbol{y}_{\boldsymbol{t}} = \frac{\boldsymbol{B}(\boldsymbol{\mathcal{B}})}{\boldsymbol{A}(\boldsymbol{\mathcal{B}})}\boldsymbol{x}_{\boldsymbol{t-k}} + \boldsymbol{n}_{\boldsymbol{t}}$$

Onde $y_t = \nabla^{\mathrm{d}} Y_{\boldsymbol{t}}$, $x_t = \nabla^{\mathrm{d}} X_{\boldsymbol{t}}$, $n_t = \nabla^{\mathrm{d}} N_{\boldsymbol{t}}$ são processos estacionários e

$$\boldsymbol{n}_{\boldsymbol{t}} = \boldsymbol{D}^{-1}(\boldsymbol{\mathcal{B}})\boldsymbol{C}(\boldsymbol{\mathcal{B}})\boldsymbol{\varepsilon}_{\boldsymbol{t}} \quad \textbf{(4.131)}$$

Assume-se que estão disponíveis $n = N - d$ pares para análise e que X_t e Y_t (x_t e y_t se $d > 0$) constituem os desvios em relação aos valores esperados.

Se os valores iniciais x_0, y_0 e ε_0 anteriores ao inicio da série são conhecidos, então, a partir dos dados é possível determinar, para qualquer escolha dos parâmetros (k, A, B, C, D), os valores de

$$\boldsymbol{\varepsilon_t = \varepsilon_t(k, A, B, C, D | x_0, y_0, \varepsilon_0)}$$

para $t = 1, 2, \ldots, n$. Na suposição da normalidade dos resíduos $\varepsilon_t's$, é possível obter uma aproximação à estimativa da máxima verosimilhança minimizando a função da soma dos quadrados condicional

$$\boldsymbol{S_0(k, A, B, C, D) = \sum_{t=1}^{n} a_t^2\ (k, A, B, C, D | x_0, y_0, a_0)} \quad \textbf{(4.132)}$$

Dados os valores iniciais apropriados, o cálculo dos $\varepsilon_t's$ podem ser calculados para qualquer escolha dos valores dos parâmetros usando um procedimento em três estágios:

1. Calcular as saídas $\hat{y}_t$ do modelo da função de transferência a partir da equação $\hat{y}_t = A^{-1}(\mathcal{B})B(\mathcal{B})x_{t-k}$ ou outro.
2. Calcular os resíduos $n_t = y_t - \hat{y}_t$
3. Os valores $\varepsilon_t's$ podem ser calculados a partir de (4.131) escrito na forma $\varepsilon_t = D(\mathcal{B})/C(\mathcal{B})n_t$

A função da soma dos quadrados é tipicamente não linear nos parâmetros, pelo que deverá ser resolvida recorrendo a metodologias de cálculo numérico, ou de forma iterativa, como discutido anteriormente para os modelos ARIMA.

Validação do modelo

Neste ponto pretende-se encontrar um procedimento que permita inferir sobre a qualidade do modelo que se obteve, ou seja, se o modelo é ou não adequado à situação onde será aplicado.

O modelo do sistema foi bem identificado se conseguir captar correctamente, e em simultâneo, as características dinâmicas e estacionárias do sistema físico, com base no conjunto de dados de estimação.

É aconselhável que se proceda, quando possível, a uma validação cruzada do modelo, isto é, um teste sobre a capacidade do modelo reproduzir um conjunto de dados diferente do utilizado durante o processo de estimação dos parâmetros, designado por conjunto de dados de teste.

Validação por análise de resíduos

Por um lado, para que o modelo possa ser considerado bom, é necessário garantir a total independência entre a entrada do sistema X_t, e os resíduos do modeloε_t, ou seja, um bom

modelo deve conseguir distinguir o que na saída depende da entrada, e o que da saída provém das perturbações.

Por outro lado, por construção do modelo, os resíduos ε_t, devem constituir uma série de ruído branco, pelo que não deve existir qualquer autocorrelação para as diversas lags $k \neq 0$.

Pode então concluir-se que para um modelo ideal, os valores dos coeficientes de correlação referidos deverão ser baixos, ou seja, devem estar dentro de um intervalo de confiança, tipicamente 2 ou 3 desvios padrão.

Esta análise poderá ser melhor interpretada com recurso à matriz dos coeficientes de correlação entre a entrada e os resíduos

$$\mathbf{r}(\boldsymbol{k}) = \begin{bmatrix} \boldsymbol{r_{xx}(k)} & \boldsymbol{r_{\varepsilon x}(k)} \\ \boldsymbol{r_{x\varepsilon}(k)} & \boldsymbol{r_{\varepsilon\varepsilon}(k)} \end{bmatrix}$$

ou, tendo em conta que se procedeu ao branqueamento das séries,

$$\mathbf{r}(\boldsymbol{k}) = \begin{bmatrix} \boldsymbol{r_{\alpha\alpha}(k)} & \boldsymbol{r_{\varepsilon\alpha}(k)} \\ \boldsymbol{r_{\alpha\varepsilon}(k)} & \boldsymbol{r_{\varepsilon\varepsilon}(k)} \end{bmatrix} \tag{4.133}$$

ou seja, pelo que foi dito acima, pode concluir-se que os valores referentes à ultima linha da matriz (4.133) devem ser estatisticamente não significativos.

Um teste global para a validação do modelo nos primeiros k coeficientes de correlação dos resíduos é dado pela estatística qui-quadrado (Del Castillo, 2002):

$$\boldsymbol{Q_1 = m(m+2)\sum_{j=1}^{k} \frac{r_{\varepsilon\varepsilon}(k)}{m-j}} \tag{4.134}$$

onde $m = n - max(r, s+k) - p$. Esta estatística, em que a hipótese nula é $H_0: \rho_{\varepsilon\varepsilon}(j) = 0$, para $j = 1, 2, \ldots, k$, segue uma distribuição χ^2_{k-q-p}. Evidentemente, a zona de rejeição é o lado direito da cauda da distribuição, pelo que valores elevados significarão elevadas autocorrelações.

De forma similar, um teste global para a estatística de teste para $H_0: \rho_{\alpha\varepsilon}(j) = 0$, para $j = 1, 2, \ldots, k$, será dada por

$$\boldsymbol{Q_2 = m(m+2)\sum_{j=1}^{k} \frac{r_{\alpha\varepsilon}(k)}{m-j}} \tag{4.135}$$

que segue também uma distribuição $\chi^2_{k+1-(r+s+1)}$.

4.3.3 Estimação recursiva

Em certas aplicações práticas torna-se necessário estimar os parâmetros do modelo on-line com recurso a algoritmos de estimação recursiva ou sequencial, como são os casos em que não existem dados disponíveis, ou estes são escassos, como por exemplo as pequenas

séries ou *short-run*. Este tipo de estimação recursiva dos parâmetros assume um papel de grande relevo no controlo adaptativo, onde os preditores e são calculados on-line.

4.3.3.1 Algoritmo recursivo dos mínimos quadrados

O objectivo passa por reescrever a solução da minimização do erro de predição quadrático (*LSE – Least Square Estimator*) de uma forma recursiva, ou seja, encontrar uma expressão genérica para a actualização do vector dos parâmetros θ_t do modelo estatístico linear

$$\boldsymbol{Y_t = Z_t\theta_t + \varepsilon_t}$$

sob a forma:

$$\boldsymbol{\widehat{\theta}_t = \widehat{\theta}_{t-1} + K_t\varepsilon_t} \quad \textbf{(4.136)}$$

onde $\boldsymbol{Y_t}$ é um vector $(tx1)$ que contem as ultima t observações, $\boldsymbol{\theta_t}$ é o vector $(nx1)$ dos parâmetros, Z_t é a matriz (txn) dos regressores, $\hat{\theta}_{t-1}$ é o vector dos parâmetros calculado no instante anterior, ε_t é o valor da nova medida do erro, ou erro de predição, ou inovação e P_t é o termo de correcção, ou ganho, que pesa o valor do erro de predição actual ε_t na estimação do novo vector de parâmetros.

No que se segue, matriz Z_t tem a seguinte estrutura

$$\boldsymbol{Z_t} = \begin{bmatrix} \boldsymbol{z_1'} \\ \boldsymbol{z_1'} \\ \vdots \\ \boldsymbol{z_1'} \end{bmatrix} = \begin{bmatrix} \boldsymbol{z_{11}} & \boldsymbol{z_{12}} & \cdots & \boldsymbol{z_{1n}} \\ \boldsymbol{z_{21}} & \boldsymbol{z_{22}} & \cdots & \boldsymbol{z_{tn}} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{z_{t1}} & \boldsymbol{z_{t2}} & \cdots & \boldsymbol{z_{tn}} \end{bmatrix}$$

Onde cada vector z_i contém o s valores dos regressores utilizados a quando da observação i.

O critério dos mínimos quadrados consiste da minimização da soma dos quadrados dos erros

$$\min_{\boldsymbol{\theta_t}} \boldsymbol{SS(\theta_t) = \varepsilon_t'\varepsilon_t = (Y_t - Z_t\theta_t)'(Y_t - Z_t\theta_t) = \frac{1}{t}\sum_{i=1}^{t}(y_i - \theta_i z_i)^2}$$

O minimizador off-line desta função é dado pelo estimador dos mínimos quadrados:

$$\boldsymbol{\widehat{\theta}_t = (Z_t'Z_t)^{-1}Z_t'Y_t} \quad \textbf{(4.137)}$$

Para a obtenção da versão recursiva desta fórmula começa-se por definir a matriz

$$\boldsymbol{R_t = \left(\sum_{i=1}^{t} \alpha_i z_i z_i'\right)} \quad \textbf{(4.138)}$$

Em que α_i é o factor de esquecimento, que se considera normalmente variante no tempo, $0 < \alpha_i \leq 1$, pesando menos o erro de predição obtido com base nas observações individuais, relativamente às observações mais recentes. Para o caso de se considerar α_i constante e igual a 1, a equação anterior fica na forma (Del Castillo, 2002)

$$\boldsymbol{R_t} = \left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{t}} \boldsymbol{z_i z_i^{'}}\right) = (\boldsymbol{Z_t^{'} Z_t}) \tag{4.139}$$

A solução genérica da minimização do erro de predição quadrático é dado por (Botto, Controlo Óptimo, 2007)

$$\boldsymbol{\widehat{\theta}_t} = (\boldsymbol{R_t})^{-1}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{t}} \boldsymbol{\alpha_i z_i\, y_i}\right) \tag{4.140}$$

Da expressão anterior pode tirar-se respectivamente para $t-1$ e t:

$$\boldsymbol{R_{t-1}\widehat{\theta}_{t-1}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{t-1}} \boldsymbol{\alpha_i z_i\, y_i}$$

$$\boldsymbol{R_t \widehat{\theta}_t} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{t}} \boldsymbol{\alpha_i z_i\, y_i} = \underbrace{\sum_{\boldsymbol{i=1}}^{\boldsymbol{t-1}} \boldsymbol{\alpha_i z_i y_i}}_{\boldsymbol{R_{t-1}\widehat{\theta}_{t-1}}} + \boldsymbol{\alpha_t z_t y_t}$$

Ou seja

$$\boldsymbol{\widehat{\theta}_t} = (\boldsymbol{R_t})^{-1}\left(\boldsymbol{R_{t-1}\widehat{\theta}_{t-1}} + \boldsymbol{\alpha_t z_t y_t}\right) \tag{4.141}$$

De forma similar, da expressão (4.138) pode tirar-se respectivamente para $t-1$ e t:

$$\boldsymbol{R_{t-1}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{t-1}} \boldsymbol{\alpha_i z_i\, z_i^{'}}$$

$$\boldsymbol{R_t} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{t}} \boldsymbol{\alpha_i z_i\, z_i^{'}} = \underbrace{\sum_{\boldsymbol{i=1}}^{\boldsymbol{t-1}} \boldsymbol{\alpha_i z_i z_i^{'}}}_{\boldsymbol{R_{t-1}}} + \boldsymbol{\alpha_t z_t z_i^{'}}$$

ou seja

$$\boldsymbol{R_t} = \boldsymbol{R_{t-1}} + \boldsymbol{\alpha_t z_t z_i^{'}} \quad \Leftrightarrow \quad \boldsymbol{R_{t-1}} = \boldsymbol{R_t} - \boldsymbol{\alpha_t z_t z_i^{'}}$$

Substituindo esta ultima expressão em (4.141) resulta finalmente:

$$\begin{aligned}\boldsymbol{\widehat{\theta}_t} &= (\boldsymbol{R_t})^{-1}\left[(\boldsymbol{R_t} - \boldsymbol{\alpha_t z_t z_t^{'}})\boldsymbol{\widehat{\theta}_{t-1}} + \boldsymbol{\alpha_t z_t y_t}\right] \\ &= (\boldsymbol{R_t})^{-1}\left(\boldsymbol{R_t}\,\boldsymbol{\widehat{\theta}_{t-1}} - \boldsymbol{\alpha_t z_t z_t^{'}}\,\boldsymbol{\widehat{\theta}_{t-1}} + \boldsymbol{\alpha_t z_t y_t}\right)\end{aligned}$$

Conseguindo-se assim uma expressão na forma (4.136), ou seja, os parâmetros recursivos dos estimadores dos mínimos quadrados podem ser determinados através do algoritmo:

$$\widehat{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\theta}}_{t-1} + \underbrace{(\boldsymbol{R}_t)^{-1}\boldsymbol{\alpha}_t \boldsymbol{z}_t}_{\boldsymbol{K}_t} \underbrace{\left(\boldsymbol{y}_t - \boldsymbol{z}_t^{'}\widehat{\boldsymbol{\theta}}_{t-1}\right)}_{\varepsilon_t} \qquad \textbf{(4.142)}$$

$$\boldsymbol{R}_t = \boldsymbol{R}_{t-1} + \boldsymbol{\alpha}_t \boldsymbol{z}_t \boldsymbol{z}_i^{'}$$

Para a implementação deste algoritmo recursivo é necessário inverter a matriz $\boldsymbol{R_t}$ em cada instante. A solução para este problema é dada pelo lema de inversão de matrizes (Del Castillo, 2002):

$$[\boldsymbol{A} + \boldsymbol{BCD}]^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{DA}^{-1}\boldsymbol{B} + \boldsymbol{C}^{-1}\right)^{-1}\boldsymbol{DA}^{-1}$$

Onde A e C são matrizes não singulares. Adaptando este lema à matriz $\boldsymbol{R_t}$ na forma

$$\boldsymbol{R}_t = \underbrace{\boldsymbol{R}_{t-1}}_{\text{A}} + \underbrace{\boldsymbol{\alpha}_t}_{\text{C}}\underbrace{\boldsymbol{z}_t}_{\text{B}}\underbrace{\boldsymbol{z}_i^{'}}_{\text{D}}$$

Resulta então:

$$\boldsymbol{R}_t^{-1} = \boldsymbol{R}_{t-1}^{-1} - \boldsymbol{R}_{t-1}^{-1}\boldsymbol{z}_t\left(\boldsymbol{z}_t^{'}\boldsymbol{R}_{t-1}^{-1}\boldsymbol{z}_t + \frac{1}{\boldsymbol{\alpha}_t}\right)^{-1}\boldsymbol{z}_t^{'}\boldsymbol{R}_{t-1}^{-1}$$

Definindo-se uma nova matriz $\boldsymbol{P_t} = \boldsymbol{R_t^{-1}}$, denominada matriz de covariância, resulta:

$$\boldsymbol{P}_t = \boldsymbol{P}_{t-1} - \boldsymbol{P}_{t-1}\boldsymbol{z}_t\left(\boldsymbol{z}_t^{'}\boldsymbol{P}_{t-1}\boldsymbol{z}_t + \frac{\boldsymbol{1}}{\boldsymbol{\alpha}_t}\right)^{-1}\boldsymbol{z}_t^{'}\boldsymbol{P}_{t-1}$$

$$= \boldsymbol{P}_{t-1} - \frac{\boldsymbol{P}_{t-1}\boldsymbol{z}_t\boldsymbol{z}_t^{'}\boldsymbol{P}_{t-1}}{\boldsymbol{z}_t^{'}\boldsymbol{P}_{t-1}\boldsymbol{z}_t + \frac{\boldsymbol{1}}{\boldsymbol{\alpha}_t}}$$

Por outro lado, do termo de correcção (ou ganho) da primeira expressão (4.142) vem

$$\boldsymbol{K}_t = (\boldsymbol{R}_t)^{-1}\boldsymbol{\alpha}_t\boldsymbol{z}_t$$

$$= \left[\boldsymbol{P}_{t-1} - \frac{\boldsymbol{P}_{t-1}\boldsymbol{z}_t\boldsymbol{z}_t^{'}\boldsymbol{P}_{t-1}}{\boldsymbol{z}_t^{'}\boldsymbol{P}_{t-1}\boldsymbol{z}_t + \frac{\boldsymbol{1}}{\boldsymbol{\alpha}_t}}\right]\boldsymbol{\alpha}_t\boldsymbol{z}_t$$

$$= \boldsymbol{P}_{t-1}\boldsymbol{\alpha}_t\boldsymbol{z}_t - \frac{\boldsymbol{P}_{t-1}\boldsymbol{z}_t\boldsymbol{z}_t^{'}\boldsymbol{P}_{t-1}\boldsymbol{\alpha}_t\boldsymbol{z}_t}{\boldsymbol{z}_t^{'}\boldsymbol{P}_{t-1}\boldsymbol{z}_t + \frac{\boldsymbol{1}}{\boldsymbol{\alpha}_t}}$$

$$= \frac{\boldsymbol{P}_{t-1}\boldsymbol{z}_t}{\boldsymbol{z}_t^{'}\boldsymbol{P}_{t-1}\boldsymbol{z}_t + \frac{\boldsymbol{1}}{\boldsymbol{\alpha}_t}}$$

Com isto pode-se finalmente estabelecer o algoritmo recursivo dos mínimos quadrados (*RLS- Recursive Least Square*). Para $t = 1, 2, 3, \ldots$, calcular sequencialmente

$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{t}} = \widehat{\boldsymbol{\theta}}_{\boldsymbol{t-1}} + \boldsymbol{K_t}\left[\boldsymbol{y_t} - \boldsymbol{z_t^{'}}\widehat{\boldsymbol{\theta}}_{\boldsymbol{t-1}}\right]$$

$$\boldsymbol{K_t} = \frac{\boldsymbol{P_{t-1}z_t}}{\boldsymbol{z_t^{'}P_{t-1}z_t} + \dfrac{\mathbf{1}}{\boldsymbol{\alpha_t}}} \qquad \textbf{(4.143)}$$

$$\boldsymbol{P_t} = \boldsymbol{P_{t-1}} - \boldsymbol{K_t z_t^{'} P_{t-1}}$$

As condições iniciais do algoritmo têm de ser definidas. (Botto, Controlo Óptimo, 2007) propõe como condições iniciais:

1. O vector dos parâmetros θ_0 considera-se normalmente igual a zero.
2. O valor inicial da matriz de covariância, P_0 considera-se normalmente $P_0 = \rho I, \rho > 0$ onde o valor a adoptar para a constante ρ influencia simultaneamente, a rapidez de convergência dos parâmetros do modelo, e a estabilidade dessa convergência (Del Castillo, 2002) propõe valores na ordem de 100 ou 1000.
3. O factor de esquecimento α_t considera-se normalmente variante no tempo, $0 < \alpha_t \leq 1$, com $\lim_{t\to\infty} \alpha_t = 1$, pesando o erro de predição obtido com base nas observações iniciais, relativamente às observações mais recentes.

4.3.3.2 Algoritmo recursivo estendido dos mínimos quadrados

Quando se pretende estimar os parâmetros de um modelo com a estrutura

$$\boldsymbol{Y_t} = \boldsymbol{Z_t\theta_t} + \boldsymbol{v_t}$$

Em que $\boldsymbol{v_t}$ não é constituído por ruído branco, o algoritmo dos mínimos quadrados recursivo fornece estimativas enviesadas. Uma situação importante em que tal acontece é nos modelos de séries temporais onde o termo média móvel (MA) aparece. Considere-se por exemplo o modelo ARMAX

$$\boldsymbol{A(\mathcal{B})Y_t} = \boldsymbol{B(\mathcal{B})X_t} + \boldsymbol{C(\mathcal{B})\varepsilon_t}$$

Onde $\{\boldsymbol{\varepsilon_t}\}$ é uma sequência de ruído branco. A expressão anterior também pode ser escrita na forma

$$\boldsymbol{Y_t} = -[\boldsymbol{A(\mathcal{B})} - \mathbf{1}]\boldsymbol{Y_t} + \boldsymbol{B(\mathcal{B})X_t} + \underbrace{\boldsymbol{C(\mathcal{B})\varepsilon_t}}_{\boldsymbol{v_t}}$$

Onde evidentemente v_t não é ruído branco se $C(\mathcal{B}) \neq \mathbf{1}$. Como se viu anteriormente, os modelos, tais como ARIMA ou ARMAX, em que o termo MA está presente, são não lineares nos parâmetros, pelo que o método dos mínimos quadrados ordinário não é aplicável, pelo que será necessário recorrer à minimização da soma dos mínimos quadrados ou função da máxima verosimilhança. Na estimação recursiva, um modo de resolver este dilema passa pelo recurso ao algoritmo dos mínimos quadrados estendido.

O algoritmo recursivo estendido dos mínimos quadrados, ou RELS (*Recursive Extended Least Squares*) corresponde à extensão do algoritmo RLS para os modelos de regressão pseudo-linear:

$$\hat{\boldsymbol{y}}_{\boldsymbol{t}}(\boldsymbol{\theta}) = \boldsymbol{\theta}^{'}\boldsymbol{z}_{\boldsymbol{t}}(\boldsymbol{\theta}) = \boldsymbol{z}_{\boldsymbol{t}}^{'}(\boldsymbol{\theta})\boldsymbol{\theta}$$

Onde o vector regressor foi construído com base em

$$\left.\begin{array}{rcl} \boldsymbol{e}_{\boldsymbol{t-1}} & = & \boldsymbol{y}_{\boldsymbol{t-1}} - \boldsymbol{\theta}^{'}\boldsymbol{z}_{\boldsymbol{t-1}} \\ \boldsymbol{e}_{\boldsymbol{t-1}} & = & \boldsymbol{y}_{\boldsymbol{t-2}} - \boldsymbol{\theta}^{'}\boldsymbol{z}_{\boldsymbol{t-2}} \\ & \vdots & \\ \boldsymbol{e}_{\boldsymbol{t-n_c}} & = & \boldsymbol{y}_{\boldsymbol{t-n_c}} - \boldsymbol{\theta}^{'}\boldsymbol{z}_{\boldsymbol{t-n_c}} \end{array}\right\} \boldsymbol{e}_{\boldsymbol{t-i}} = \boldsymbol{y}_{\boldsymbol{t-i}} - \boldsymbol{\theta}^{'}\boldsymbol{z}_{\boldsymbol{t-i}}$$

Dado o carácter iterativo da estimação do vector de parâmetros, não é possível conhecer com exactidão, num determinado instante t, o vector de parâmetros do modelo, θ. Para ultrapassar este problema, considera-se sempre o valor de $\hat{\theta}_{t-1}$ nas aproximações dos termos $e_{t-1}, \dots, e_{t-n_c}$ presentes no vector regressor:

$$\left.\begin{array}{rcl} \boldsymbol{e}_{\boldsymbol{t-1}} & \approx & \boldsymbol{y}_{\boldsymbol{t-1}} - \boldsymbol{\theta}_{\boldsymbol{t-1}}^{'}\boldsymbol{z}_{\boldsymbol{t-1}} \\ \boldsymbol{e}_{\boldsymbol{t-1}} & \approx & \boldsymbol{y}_{\boldsymbol{t-2}} - \boldsymbol{\theta}_{\boldsymbol{t-1}}^{'}\boldsymbol{z}_{\boldsymbol{t-2}} \\ & \vdots & \\ \boldsymbol{e}_{\boldsymbol{t-n_c}} & \approx & \boldsymbol{y}_{\boldsymbol{t-n_c}} - \boldsymbol{\theta}_{\boldsymbol{t-1}}^{'}\boldsymbol{z}_{\boldsymbol{t-n_c}} \end{array}\right\} \boldsymbol{e}_{\boldsymbol{t-i}} \approx \boldsymbol{y}_{\boldsymbol{t-i}} - \boldsymbol{\theta}_{\boldsymbol{t-1}}^{'}\boldsymbol{z}_{\boldsymbol{t-i}}$$

O algoritmo recursivo estendido dos mínimos quadrados, ou algoritmo RELS (*Recursive Extended Least Square*) vem dado por

$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{t}} = \widehat{\boldsymbol{\theta}}_{\boldsymbol{t-1}} + \boldsymbol{K}_{\boldsymbol{t}}\left[\boldsymbol{y}_{\boldsymbol{t}} - \boldsymbol{z}_{\boldsymbol{t}}^{'}\widehat{\boldsymbol{\theta}}_{\boldsymbol{t-1}}\right]$$

$$\boldsymbol{K}_{\boldsymbol{t}} = \frac{\boldsymbol{P}_{\boldsymbol{t-1}}\boldsymbol{z}_{\boldsymbol{t}}}{\boldsymbol{z}_{\boldsymbol{t}}^{'}\boldsymbol{P}_{\boldsymbol{t-1}}\boldsymbol{z}_{\boldsymbol{t}} + \dfrac{\boldsymbol{1}}{\boldsymbol{\alpha}_{\boldsymbol{t}}}} \qquad \textbf{(4.144)}$$

$$\boldsymbol{P}_{\boldsymbol{t}} = \boldsymbol{P}_{\boldsymbol{t-1}} - \boldsymbol{K}_{\boldsymbol{t}}\boldsymbol{z}_{\boldsymbol{t}}^{'}\boldsymbol{P}_{\boldsymbol{t-1}}$$

$$\boldsymbol{e}_{\boldsymbol{t-i}} = \boldsymbol{y}_{\boldsymbol{t-i}} - \boldsymbol{\theta}_{\boldsymbol{t-1}}^{'}\boldsymbol{z}_{\boldsymbol{t-i}}, \qquad i = 1, \dots, n_c$$

5 Sistemas de Controlo

Os sistemas de controlo comuns podem ser classificados *sistemas de controlo de referência* ou *sistemas de controlo terminal*. Nos sistemas de controlo de referência, a variável controlada $y(t)$ tem de seguir uma referência variável $w(t)$ tão próximo quanto possível, resultando em erros de controlo $e(t) = w(t) - y(t)$ que devem ser tão pequenos quanto possível, $e(t) \approx 0$. Se a variável de referência muda com o tempo está-se perante um *sistema de controlo de referência variável.* Se a variável de referência é constante, então este denomina-se como um *regulador*.

Para sistemas de controlo terminal, o objectivo é atingir-se e manter-se um estado final do processo ao fim de um determinado instante N (predito ou livre). Tanto para os sistemas de controlo de referência como para os sistemas de controlo terminal, os valores iniciais ou as perturbações dos processos têm de ser compensadas o máximo possível. Além disso, o problema de controlo passa por tornar processos instáveis em sistemas globais estáveis conseguidos através de controlo por realimentação ou *feedback*.

Estes problemas podem ser resolvidos, em geral pela aplicação de controladores que utilizam a informação retirada das saídas ou estados dos processos para realimentar o próprio processo. O efeito de *feedback* pode frequentemente ser melhorado com a introdução de elementos controlo de compensação em avanço (*feedforward*).

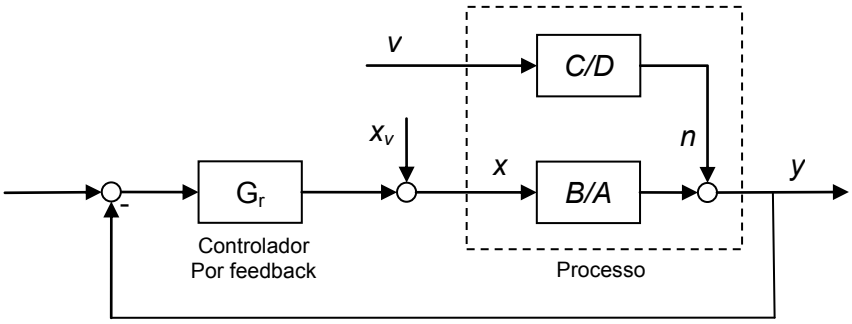


Figura 5-1 Diagrama de blocos de controlo em anel fechado

A Figura 5-1 mostra um anel de controlo simples. Se as perturbações *v* são mensuráveis, pode-se utilizar a compensação em avanço (feedforward) como na Figura 5-2, em combinação com o anel de realimentação (*feddback loop*) para controlo dos distúrbios que não podem ser compensados pelo controlo em avanço.

O projecto de controlo de sistemas é desenvolvido de acordo com o gráfico da Figura 5-3.Dependendo do método de projecto e da aplicação, são utilizados como base de projecto os modelos matemáticos dos processos e dos sinais (distúrbios, variáveis de referencia, valores iniciais). Frequentemente, os modelos dos sinais apenas se conseguem estimar de forma aproximada. Por simplicidade, assume-se frequentemente alterações em degrau, apesar de serem raras na prática. Contudo, com recurso à capacidade actual de

computação, é possível obter-se modelos mais exactos de sinais determinísticos e estocásticos sem grandes esforços.

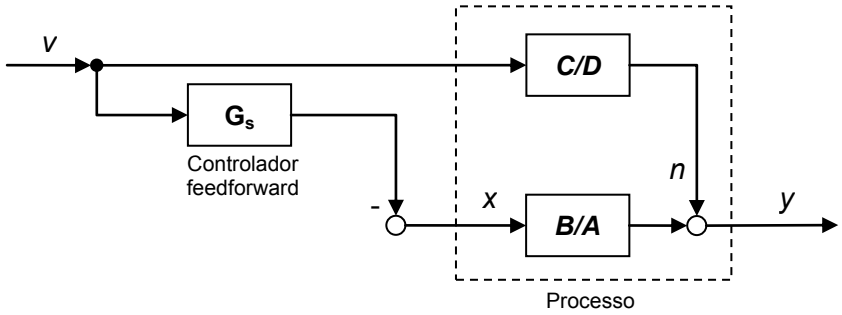


Figura 5-2 Diagrama de blocos de sistema de controlo de alimentação em avanço (feedforward)

No projecto de controladores lineares há que distinguir dois tipos de sistemas de controlo: sistemas de controlo de parâmetros optimizados e sistemas de controlo de estrutura optimizada.

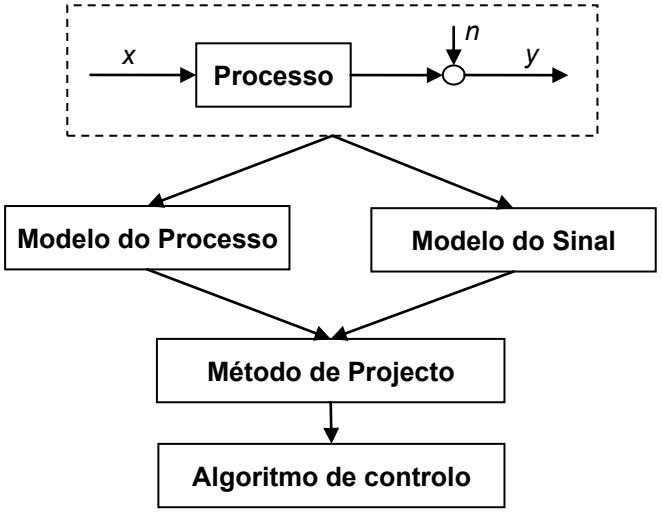


Figura 5-3 Metodologia de projecto de algoritmos de controlo (Isermann, 1989)

No caso de sistemas de controlo de parâmetros optimizados, a estrutura do controlador , isto é, a forma e a ordem da equação do controlador, é dada e os parâmetros livres do controlador são adaptados ao processo controlado através de critérios de optimização ou regras de sintonização. Sistemas de controlo de estrutura optimizada são sistema em que a estrutura e os parâmetros do controlador são adaptados e optimizados à estrutura e parâmetros do modelo do processo. No primeiro caso tem-se os controladores tipo PID, enquanto que no segundo caso pode enunciar-se os controladores de estado e os controladores de cancelamento (Isermann, 1989). Para projecto, pode recorrer-se a regras de sintonização, critérios de performance e colocação de pólos.

A escolha dos critérios de performance tem um papel central no projecto de controladores. Para controlo de sistemas contínuos, o critério integral tem sido dos mais utilizados, pela integração dos erros, erros quadráticos, amplitude absoluta, etc., cada um podendo ser

ponderados no tempo. Para sinais discretos, os critérios utilizados tem sido os seguintes (Isermann, 1989):

$$\boldsymbol{I_1 = \sum_{t=0}^{\infty} e_t} \qquad \textbf{“Soma do erro”}$$

$$\boldsymbol{I_2 = \sum_{t=0}^{\infty} e_t^2} \qquad \textbf{“Soma do quadrado do erro”}$$

$$\boldsymbol{I_3 = \sum_{t=0}^{\infty} |e_t|} \qquad \textbf{“Soma da amplitude absoluta do erro”}$$

$$\boldsymbol{I_4 = \sum_{t=0}^{\infty} t|e_t|} \qquad \textbf{“Soma de tempo multiplicado pela amplitude absoluta do erro”}$$

A utilização de $\boldsymbol{I_1}$ não é aconselhável se $\boldsymbol{e_t}$ mudar de sinal, pelo que $\boldsymbol{I_2}$ é utilizado mais frequentemente. Contudo, $\boldsymbol{I_2}$ conduz a um comportamento fortemente oscilatório ao passo que, com $\boldsymbol{I_3}$ e $\boldsymbol{I_4}$ obtém-se um comportamento mais amortecido das variáveis de controlo.

No projecto analítico prefere-se os critérios de performance quadráticos devido às suas vantagens matemáticas. Isso deve-se ao facto de que quando se procura valores extremos, uma simples derivada resulta em relações que são lineares em $\boldsymbol{e_t}$. Se aos desvios quadráticos das variáveis controladas se adicionar os desvios quadráticos das variáveis manipuladas, ponderadas com um factor r, obtém-se um acrescento dos graus de liberdade e a possibilidade de uma mais directa infuencia no amortecimento do comportamentodo sistema de controlo. Pelo que, um critério de performance quadrática mais geral será dado pela expressão:

$$\boldsymbol{I_5 = \sum_{t=0}^{\infty} \left[e_t^2 + r x_t^2\right]}$$

Que, para sistema de controlo de estado conduz à forma

$$\boldsymbol{I_6 = \sum_{t=0}^{\infty} \left[y_t^T Q y_t + r x_t^2\right]}$$

Estes critérios quadráticos são adequados tanto para sinais determinísticos como estocásticos (Isermann, 1989).

5.1 Controlo óptimo

Nesta secção pretende-se resolver o seguinte problema: como ajustar o processo com a garantia de que os ajustamentos efectuados são os mais adequados para que o erro da saída do processo $e(t)$ seja o mínimo possível. Para isso apresenta-se uma derivação geral de, como se obter um controlador de variância mínima ou MMSE (*Minimum Mean Square Error*). Apesar de utópicos, ou seja, geralmente (fisicamente) impraticáveis, os controladores MMSE fornecem uma intuição considerável e são facilmente modificáveis a fim de se obterem alternativas robustas mais realisticamente implementáveis.

5.1.1 Controladores de variância mínima (MMSE)

Estratégias de controlo que envolvam apenas a minimização da variável de saída, ignorando todos os outros custos envolvidos, tornam-se normalmente ineficazes, devido não só aos esforços envolvidos nas variáveis manipuladas, como também à dificuldade da sua implementação física. No entanto, o modelo obtido através desta estratégia poderá conduzir a estratégias de controlo bastante eficazes tanto no campo da implementação física como económica, devido à introdução de funções de custo que envolvam para além da minimização da variância da variável de saída, outros factores a ter em conta, como sejam os custos e as condições operacionais das variáveis de entrada: variabilidade, amplitude, tempo de resposta etc. Adicionalmente, a derivação da estratégia de controlo de variância mínima fornece muita informação sobre os problemas de ajustamento inerentes ao processo. Para se ter uma ideia mais clara sobre este assunto tome-se como exemplo o processo ARX descrito por (Del Castillo, 2002):

$$\boldsymbol{Y_t = \mu + bX_{t-1} + \varepsilon_t}$$

Onde o μ é a média do processo se não existir qualquer ajustamento na variável manipulada, ou seja, se $X_t = \mathbf{0}$. Se o objectivo é obter-se o mínimo desvio quadrático da variável de saída e simultaneamente que esta mantenha o valor desejado, ou seja, o valor de referência ou *target* T, então o objectivo será $E[Y_t] = T$, que se consegue fazendo

$$\boldsymbol{E[\mu + bX_t + \varepsilon_t] = T} \quad \Rightarrow \quad \boldsymbol{X_t = \frac{T - \mu}{b}}$$

Substituindo este resultado na expressão acima obtém que $\boldsymbol{Y_t = T} + \varepsilon_t$ pelo que o erro mínimo será ε_t cuja variância corresponderá à variância mínima $\boldsymbol{\sigma_\varepsilon^2}$. Isto implica que os ajustamentos ∇X_t serão sempre iguais a zero. Se $\boldsymbol{X}\_0 \neq (\boldsymbol{T - \mu})/\boldsymbol{b}$, será necessário apenas um ajustamento. Se $\mu = T$, não será necessário qualquer ajustamento. Este resultado está de acordo com os pressupostos da experiência do funil de Demming, onde apenas se o funil está fora do alvo (*target)*, este apenas necessita de ser ajustado ao alvo, após o que não será necessário mais nenhum ajustamento. Este modelo de processos apenas necessitam de ser monitorizados para se garantir que não há qualquer alteração, pelo que a utilização de técnicas de controlo estatístico serão suficientes para os objectivos pretendidos.

5.1.1.1 Predição através da função de transferência

Suponha-se que a função de transferência de um determinado processo é dada pelo modelo geral

$$\boldsymbol{Y_t = \frac{B_n(\mathcal{B})}{A_n(\mathcal{B})} X_{t-k}}$$

Que também pode ser escrito na forma

$$\boldsymbol{y_t = -a_1 y_{t-1} - \cdots - a_n y_{t-n} + b_0 x_{t-k} + \cdots + b_n x_{t-k-n}} \qquad \textbf{(5.1)}$$

Se o grau dos dois polinómios não for o mesmo, é sempre possível substituir os coeficientes inexistentes por zero. Pode-se então utilizar este modelo para a base para predição, se

$d \geq 1$. Aqui assume-se que o valor de $\boldsymbol{x_t}$ não é ainda conhecido quando se faz a predição da variável de saída i instantes à frente $(\hat{y}_{t+i|t})$.

A maneira óbvia de se obter a predição da saída é como se segue

$$\hat{\boldsymbol{y}}_{\boldsymbol{t+1|t}} = -\sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} \boldsymbol{a_j y_{t+1-j}} + \sum_{\boldsymbol{j=0}}^{\boldsymbol{n}} \boldsymbol{b_j \breve{x}_{t-k-j}}$$

$$\hat{\boldsymbol{y}}_{\boldsymbol{t+2|t}} = -\boldsymbol{a_j \hat{y}_{t+1|t}} - \sum_{\boldsymbol{j=2}}^{\boldsymbol{n}} \boldsymbol{a_j y_{t+2-j}} + \sum_{\boldsymbol{j=0}}^{\boldsymbol{n}} \boldsymbol{b_j \breve{x}_{t+1-k-j}}$$

ou, em geral

$$\hat{\boldsymbol{y}}_{\boldsymbol{t+i|t}} = -\sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} \boldsymbol{a_j \breve{y}_{t+1-j}} + \sum_{\boldsymbol{j=0}}^{\boldsymbol{n}} \boldsymbol{b_j \breve{x}_{t-k-j}}$$

ou

$$\boldsymbol{A(\mathcal{B})\breve{y}_{t+i} = \mathcal{B}^k B(\mathcal{B})\breve{x}_{t-i}} \quad \textbf{(5.2)}$$

onde

$$\breve{\boldsymbol{x}}_{\boldsymbol{l}} = \begin{cases} \boldsymbol{x_l} & \boldsymbol{se\ l \leq t} \\ \boldsymbol{\hat{x}_{l|t}} & \boldsymbol{se\ l < t} \end{cases}$$

e

$$\breve{\boldsymbol{y}}_{\boldsymbol{l}} = \begin{cases} \boldsymbol{y_l} & \boldsymbol{se\ l \leq t} \\ \boldsymbol{\hat{y}_{l|t}} & \boldsymbol{se\ l < t} \end{cases}$$

Desde que a predição $\hat{y}_{t+i|t}$ depende de outras saídas preditas, as quais elas próprias foram preditas em função de medidas provenientes de saídas passadas, é possível encontrar uma expressão para $\hat{y}_{t+i|t}$ que dependa apenas das medidas das saídas medidas $y_t, y_{t-1}, \ldots$ (Maciejowski, 2002).

Suponha-se que se tem um polinómio $E_i(\mathcal{B})$, de grau menor ou igual a $i-1$, e um polinómio $F_i(\mathcal{B})$, de grau $n-1$, tal que:

$$\frac{\mathbf{1}}{\boldsymbol{A(\mathcal{B})}} = \boldsymbol{E_i(\mathcal{B})} + \frac{\boldsymbol{F_i(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}}\boldsymbol{\mathcal{B}^i}$$

ou

$$\boldsymbol{E_i(\mathcal{B})A(\mathcal{B}) = 1 - F_i(\mathcal{B})\mathcal{B}^i} \quad \textbf{(5.3)}$$

Multiplicando (5.2) por $E_i(\mathcal{B})$ obtém-se

$$\left[\mathbf{1} - \boldsymbol{F_i}(\mathcal{B})\mathcal{B}^i\right]\breve{\boldsymbol{y}}_{t+i} = \mathcal{B}^k \boldsymbol{E_i}(\mathcal{B})\boldsymbol{B}(\mathcal{B})\breve{\boldsymbol{x}}_{t-i}$$

ou

$$\breve{\boldsymbol{y}}_{t+i} = \mathcal{B}^i \boldsymbol{F_i}(\mathcal{B})\breve{\boldsymbol{y}}_{t+i} + \mathcal{B}^k \boldsymbol{E_i}(\mathcal{B})\boldsymbol{B}(\mathcal{B})\breve{\boldsymbol{x}}_{t-i}$$

Deste modo conseguiu-se construir uma expressão em predição i instantes à frente da saída do processo $\breve{y}_{t+i}$ apenas depende dos valores conhecidos (presente e passados) da saída do mesmo processo (o ultimo valor possível envolvido poderá ser no máximo $\mathcal{B}^i\breve{y}_{t+i} = \breve{y}_t = y_t$). Poderá então escrever-se a predição para a saída do processo i instantes à frente na forma

$$\hat{\boldsymbol{y}}_{t+\mathrm{i}|t} = \boldsymbol{F_i}(\mathcal{B})\breve{\boldsymbol{y}}_t + \mathcal{B}^k \boldsymbol{E_i}(\mathcal{B})\boldsymbol{B}(\mathcal{B})\breve{\boldsymbol{x}}_{t-i}$$

a qual apenas envolve saídas do processo conhecidas no lado direito da expressão. Claro que tudo isto depende da existência das soluções $\boldsymbol{E_i}(\mathcal{B})$ e $\boldsymbol{F_i}(\mathcal{B})$ para a equação (5.3). Esta metodologia é um exemplo de aplicação de uma equação de *Diophantine* (Maciejowski, 2002, p. 126).

Identidade de *Diophantine*

Na presença de uma expressão do tipo

$$\boldsymbol{C}(\mathcal{B}) = \frac{\boldsymbol{B}(\mathcal{B})}{\boldsymbol{A}(\mathcal{B})}$$

em que os polinómios A e B são ambos de grau n. Se não for o caso, os respectivos coeficientes serão iguais a zero. A identidade de *Diophantine* (ou equação de *Diophantine*) representa a longa divisão de B por A, que resulta num quociente E_i e um resto $\mathcal{B}^i\, F_i/A$

$$\boldsymbol{C}(\mathcal{B}) = \boldsymbol{E_i}(\mathcal{B}) + \frac{\boldsymbol{F_i}(\mathcal{B})}{\boldsymbol{A}(\mathcal{B})}\mathcal{B}^i$$

onde $E_i(\mathcal{B})$ é um polinómio de grau menor ou igual a $i-1$, e $F_i(\mathcal{B})$ é um polinómio de grau não superior a $n-1$.

5.1.1.2 Predição com modelo de distúrbios

Neste ponto assume-se que o processo é modulado por função de transferência-ruido Box-Jenkins que poderá ser escrita na forma:

$$\boldsymbol{y_t} = \frac{\boldsymbol{B}(\mathcal{B})}{\boldsymbol{A}(\mathcal{B})}\boldsymbol{x_{t-k}} + \boldsymbol{n_t} \qquad \textbf{(5.4)}$$

$$\boldsymbol{n_t} = \frac{\boldsymbol{C}(\mathcal{B})}{\boldsymbol{D}(\mathcal{B})}\boldsymbol{\varepsilon_t} \qquad \textbf{(5.5)}$$

Na qual assume-se, como anteriormente que os polinómios $C(\mathcal{B})$ e $D(\mathcal{B})$ são do mesmo grau:

$$\boldsymbol{C}(\mathcal{B}) = \mathbf{1} + \boldsymbol{c_1}\mathcal{B} + \cdots + \boldsymbol{c_v}\mathcal{B}^v$$

$$\boldsymbol{D}(\mathcal{B}) = \mathbf{1} + \boldsymbol{d_1}\mathcal{B} + \cdots + \boldsymbol{d_v}\mathcal{B}^v$$

Onde, como anteriormente, $\boldsymbol{\varepsilon_t}$ é uma modulado como sendo ruído branco.

Quando o modelo dos distúrbios da forma (5.5) está presente, pode-se recorrer novamente à solução de uma equação de *Diophantine* para se determinar a predição da saída do processo. Nesta situação assume-se que se tem polinómios $E_i^{'}(\mathcal{B})$ e $F_i^{'}(\mathcal{B})$ que resolvem a equação

$$\frac{\boldsymbol{C(\mathcal{B})}}{\boldsymbol{D(\mathcal{B})}} = \boldsymbol{E_i^{'}(\mathcal{B})} + \frac{\boldsymbol{F_i^{'}(\mathcal{B})}}{\boldsymbol{D(\mathcal{B})}}\boldsymbol{\mathcal{B}^i}$$

em que o grau do polinómio $E_i^{'}(\mathcal{B})$ é no máximo $i-1$ e o grau do polinómio $F_i^{'}(\mathcal{B})$ é no máximo $v-1$. Ou seja, estes polinómios resolvem a equação de *Diophantine*

$$\boldsymbol{E_i^{'}(\mathcal{B})D(\mathcal{B})} = \boldsymbol{C(\mathcal{B})} - \boldsymbol{F_i^{'}(\mathcal{B})\mathcal{B}^i}$$

Note-se que, neste caso, o primeiro termo do lado direito é o polinómio $C(\mathcal{B})$, ao contrário do caso anterior que era 1. Esta diferença surge devido ao facto de a entrada do processo x ser conhecida, ao contrário do distúrbio v que é deconhecido.

Dividindo ambos os membros da equação anterior por $D(\mathcal{B})$ tem-se que

$$\frac{\boldsymbol{C(\mathcal{B})}}{\boldsymbol{D(\mathcal{B})}} = \boldsymbol{E_i^{'}(\mathcal{B})} + \frac{\boldsymbol{F_i^{'}(\mathcal{B})}}{\boldsymbol{D(\mathcal{B})}}\boldsymbol{\mathcal{B}^i}$$

Usando este resultado e (5.5), obtém-se

$$\begin{aligned}\boldsymbol{\hat{n}_{t+i|t}} &= \left[\boldsymbol{E_i^{'}(\mathcal{B})} + \frac{\boldsymbol{F_i^{'}(\mathcal{B})}}{\boldsymbol{D(\mathcal{B})}}\boldsymbol{\mathcal{B}^i}\right]\boldsymbol{\varepsilon_{t+i|t}} \\ &= \underbrace{\boldsymbol{E_i^{'}(\mathcal{B})\varepsilon_{t+i|t}}}_{\boldsymbol{futuro}} + \underbrace{\frac{\boldsymbol{F_i^{'}(\mathcal{B})}}{\boldsymbol{D(\mathcal{B})}}\boldsymbol{\varepsilon_{t|t}}}_{\boldsymbol{passado\ e\ presente}}\end{aligned} \quad \textbf{(5.6)}$$

A diferença entre a situação anterior e esta na utilização da equação de *Diophantine*, é que, apesar de se ter dividido a predição entre termos "passados" e termos "futuros", neste caso não se conhece o termo presente dos distúrbios ε_t. Esse termo terá que ser estimado de alguma forma.

Uma das formas possíveis passa por pegar no sistema formado pelas equações (5.4) e (5.5) e resolvê-lo em ordem a ε_t:

$$\begin{aligned}\boldsymbol{\hat{\varepsilon}_{t|t}} &= \frac{\boldsymbol{D(\mathcal{B})}}{\boldsymbol{C(\mathcal{B})}}\left[\boldsymbol{y_t} - \frac{\boldsymbol{B(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}}\boldsymbol{\mathcal{B}^k x_t}\right] \\ &= \frac{\boldsymbol{D(\mathcal{B})}}{\boldsymbol{C(\mathcal{B})}}[\boldsymbol{y_t} - \boldsymbol{\hat{y}_t}]\end{aligned} \quad \textbf{(5.7)}$$

onde $\hat{y}_t$ é a predição da saída obtida pela filtragem da entrada x pelo modelo do processo. O diagrama de blocos da Figura 5-4 mostra uma interpretação da expressão anterior (Maciejowski, 2002).

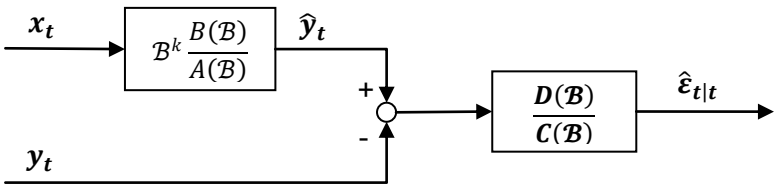


Figura 5-4 Geração de $\hat{\varepsilon}_{t|t}$ - uma interpretação (Maciejowski, 2002)

A partir da expressões (5.4) e (5.6) pode-se então construir uma expressão para a predição para a saída do processo i-instantes à frente:

$$\hat{\boldsymbol{y}}_{\boldsymbol{t+i|t}} = \frac{\boldsymbol{B}(\boldsymbol{\mathcal{B}})}{\boldsymbol{A}(\boldsymbol{\mathcal{B}})}\boldsymbol{\mathcal{B}^k}\breve{\boldsymbol{x}}_{\boldsymbol{t+i}} + \boldsymbol{E_i'}(\boldsymbol{\mathcal{B}})\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{t+i|t}} + \frac{\boldsymbol{F_i'}(\boldsymbol{\mathcal{B}})}{\boldsymbol{D}(\boldsymbol{\mathcal{B}})}\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{t|t}}$$

Inserindo (5.7) na expressão anterior obtém-se

$$\hat{\boldsymbol{y}}_{\boldsymbol{t+i|t}} = \frac{\boldsymbol{B}(\boldsymbol{\mathcal{B}})}{\boldsymbol{A}(\boldsymbol{\mathcal{B}})}\boldsymbol{\mathcal{B}^k}\breve{\boldsymbol{x}}_{\boldsymbol{t+i}} + \boldsymbol{E_i'}(\boldsymbol{\mathcal{B}})\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{t+i|t}} + \frac{\boldsymbol{F_i'}(\boldsymbol{\mathcal{B}})}{\boldsymbol{C}(\boldsymbol{\mathcal{B}})}\left[\boldsymbol{y_t} - \frac{\boldsymbol{B}(\boldsymbol{\mathcal{B}})}{\boldsymbol{A}(\boldsymbol{\mathcal{B}})}\boldsymbol{\mathcal{B}^k}\boldsymbol{x_t}\right]$$

Tal como em (5.6), esta predição inclui o termo $E_i'(\mathcal{B})\hat{\varepsilon}_{t+i|t}$, que contém apenas predições de valores futuros dos distúrbios ε_t. A melhor maneira de fazer essas previsões seria supor uma determinada natureza de distúrbios. No entanto, não seria de bom senso assumir que ε_t seja um sinal predizível de alguma forma porque, se fosse o caso, essa situação deveria reflectir-se nos polinómios C e D. Como o modelo assume que se trata de ruído branco, então os $\varepsilon_t's$ são independentes e identicamente distribuídos com valor esperado nulo, pelo que, nesta situação, faz todo o sentido fazer $\hat{\varepsilon}_{t+i} = 0$. Deste modo obtém-se finalmente o *preditor de variância mínima*:

$$\hat{\boldsymbol{y}}_{\boldsymbol{t+i|t}} = \frac{\boldsymbol{B}(\boldsymbol{\mathcal{B}})}{\boldsymbol{A}(\boldsymbol{\mathcal{B}})}\boldsymbol{\mathcal{B}^k}\breve{\boldsymbol{x}}_{\boldsymbol{t+i}} + \frac{\boldsymbol{F_i'}(\boldsymbol{\mathcal{B}})}{\boldsymbol{C}(\boldsymbol{\mathcal{B}})}\left[\boldsymbol{y_t} - \frac{\boldsymbol{B}(\boldsymbol{\mathcal{B}})}{\boldsymbol{A}(\boldsymbol{\mathcal{B}})}\boldsymbol{\mathcal{B}^k}\boldsymbol{x_t}\right] \quad \textbf{(5.8)}$$

Para utilizar esta expressão numa estratégia de controlo preditivo, há ainda a necessidade de separar a parte de 'resposta livre', a resposta predita que ocorre se não houver alteração na entrada do processo ($\Delta\breve{\boldsymbol{x}}_{\boldsymbol{t+i|t}} = 0$), da parte da "resposta forçada", nomeadamente que depende de ($\Delta\breve{\boldsymbol{x}}_{\boldsymbol{t+i|t}}$. Por outras palavras, é necessário separar a parte que depende das entradas passadas da parte que depende de entradas futuras. Para se atingir esse objectivo, recorre-se a outra equação de *Diophantine*.

Da expressão (5.8), verifica-se que o único sinal que atravessa a fronteira passado/futuro é $\breve{\boldsymbol{x}}_{\boldsymbol{t}}$, que é filtrado pela função de transferência $\boldsymbol{B}(\boldsymbol{\mathcal{B}})/\boldsymbol{A}(\boldsymbol{\mathcal{B}})$ $\boldsymbol{\mathcal{B}\text{^}k}$, pelo que se torna necessário extrair os primeiros $i-k+1$ termos da resposta ao impulso de $\boldsymbol{B}(\boldsymbol{\mathcal{B}})/\boldsymbol{A}(\boldsymbol{\mathcal{B}})$. Pelo que se necessita de determinar o par solução $\left(\boldsymbol{E_i}(\boldsymbol{\mathcal{B}}),\ \boldsymbol{F_i}(\boldsymbol{\mathcal{B}})\right)$ da equação de *Diophantine*

$$\frac{\boldsymbol{B(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}} = \boldsymbol{E_i(\mathcal{B})} + \frac{\boldsymbol{F_i(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}}\boldsymbol{\mathcal{B}^{i-k}}$$

Em que o grau de E_i não superior a $i - d$. Substituindo esta expressão em (5.8), obtém-se finalmente (Maciejowski, 2002)

$$\begin{aligned}
\widehat{\boldsymbol{y}}_{\boldsymbol{t+i|t}} &= \left[\boldsymbol{E_i(\mathcal{B})} + \frac{\boldsymbol{F_i(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}}\boldsymbol{\mathcal{B}^{i-k+1}}\right]\boldsymbol{\mathcal{B}^k}\widehat{\boldsymbol{x}}_{\boldsymbol{t+i}} + \frac{\boldsymbol{F_i'(\mathcal{B})}}{\boldsymbol{C(\mathcal{B})}}\left[\boldsymbol{y_t} - \frac{\boldsymbol{B(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}}\boldsymbol{\mathcal{B}^k x_t}\right] \\
&= \underbrace{\boldsymbol{E_i(\mathcal{B})}\widehat{\boldsymbol{x}}_{\boldsymbol{t+i-k}}}_{\boldsymbol{futuro}} + \underbrace{\frac{\boldsymbol{F_i(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}}\boldsymbol{x_{t-1}} + \frac{\boldsymbol{F_i'(\mathcal{B})}}{\boldsymbol{C(\mathcal{B})}}\left[\boldsymbol{y_t} - \frac{\boldsymbol{B(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}}\boldsymbol{\mathcal{B}^k x_t}\right]}_{\boldsymbol{passado}} \\
&= \underbrace{\boldsymbol{E_i(\mathcal{B})}\widehat{\boldsymbol{x}}_{\boldsymbol{t+i-k}}}_{\boldsymbol{futuro}} + \underbrace{\frac{\boldsymbol{F_i'(\mathcal{B})}}{\boldsymbol{C(\mathcal{B})}}\boldsymbol{y_t} + \frac{\boldsymbol{F_i(\mathcal{B})C(\mathcal{B})\mathcal{B} - F_i'(\mathcal{B})B(\mathcal{B})\mathcal{B}^k}}{\boldsymbol{A(\mathcal{B})C(\mathcal{B})}}\boldsymbol{x_t}}_{\boldsymbol{passado}}
\end{aligned} \tag{5.9}$$

5.1.1.3 Solução geral para o problema de controlo de variância mínima (MMSE)

Fazendo algumas simplificações na metodologia apresentada no ponto anterior (5.1.1.1), (Del Castillo, 2002) apresenta a solução para o problema de controlo de variância mínima.

Uma das simplificações consiste em utilizar-se um modelo ARMAX em vez do modelo Box-Jenkins. Pode-se no entanto considerar que não se trata de uma simplificação, uma vez que é possível passar-se facilmente de um modelo BJ para um modelo ARMAX:

$$\boldsymbol{Y_t} = \frac{\boldsymbol{B_1(\mathcal{B})}}{\boldsymbol{A_1(\mathcal{B})}}\boldsymbol{X_{t-k}} + \frac{\boldsymbol{C_1(\mathcal{B})}}{\boldsymbol{D(\mathcal{B})}}\boldsymbol{\varepsilon_t}$$

$$\Leftrightarrow \quad \underbrace{\boldsymbol{A_1(\mathcal{B})D(\mathcal{B})}}_{\boldsymbol{A(\mathcal{B})}}\boldsymbol{Y_t} = \underbrace{\boldsymbol{B_1(\mathcal{B})D(\mathcal{B})}}_{\boldsymbol{B(\mathcal{B})}}\boldsymbol{X_{t-k}} + \underbrace{\boldsymbol{C_1(\mathcal{B})A_1(\mathcal{B})}}_{\boldsymbol{C(\mathcal{B})}}\boldsymbol{\varepsilon_t}$$

Outra das simplificações foi fazer o horizonte de predição k coincidir com o atraso da resposta do sistema.

Considere-se então uma função de transferência ARMAX / BJ, e suponha-se que no instante t tem que se tomar uma decisão sobre o valor do ajustamento a fazer da variável de entrada:

$$\boldsymbol{Y_{t+k}} = \frac{\boldsymbol{B(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}}\boldsymbol{X_t} + \frac{\boldsymbol{C(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}}\boldsymbol{\varepsilon_{t+k}} \tag{5.10}$$

onde k é o atraso da resposta do sistema. O segundo termo do lado é composto por desvios futuros desconhecidos, $(\varepsilon_{t+1},\ \ldots\ ,\ \varepsilon_{t+k})$, e desvios passados que podem ser estimados a partir dos resíduos $(\varepsilon_t,\ \varepsilon_{t-1},\ \ldots)$. Para fazer a separação entre desvios passados e futuros, recorre-se novamente à equação de *Diophantine*, fazendo:

$$\frac{\boldsymbol{C(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}} = \boldsymbol{E(\mathcal{B})} + \frac{\boldsymbol{F(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}}\boldsymbol{\mathcal{B}^k} \quad \textbf{(5.11)}$$

pelo que a expressão (5.10) ficará então na forma

$$\boldsymbol{Y_{t+k}} = \frac{\boldsymbol{B(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}}\boldsymbol{X_t} + \boldsymbol{E(\mathcal{B})\varepsilon_{t+k}} + \frac{\boldsymbol{F(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}}\boldsymbol{\varepsilon_t} \quad \textbf{(5.12)}$$

onde, como anteriormente, E é de ordem $k-1$ e F é de ordem $n-1$, em que $n = max\{n_A, n_B, n_C\}$, onde os coeficientes inexistentes são substituídos por zero.

Para estimar os desvios passados e presente, recorre-se novamente ao modelo original (ARMAX/BJ) resolvendo em ordem a ε_t para o instante t

$$\boldsymbol{\varepsilon_t} = \frac{\boldsymbol{A(\mathcal{B})}}{\boldsymbol{C(\mathcal{B})}}\boldsymbol{Y_t} - \frac{\boldsymbol{B(\mathcal{B})}}{\boldsymbol{C(\mathcal{B})}}\boldsymbol{X_{t-k}}$$

Substituindo em (5.12), obtém-se

$$\begin{aligned}\boldsymbol{Y_{t+k}} &= \frac{\boldsymbol{B(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}}\boldsymbol{X_t} + \boldsymbol{E(\mathcal{B})\varepsilon_{t+k}} + \frac{\boldsymbol{F(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}}\left[\frac{\boldsymbol{A(\mathcal{B})}}{\boldsymbol{C(\mathcal{B})}}\boldsymbol{Y_t} - \frac{\boldsymbol{B(\mathcal{B})}}{\boldsymbol{C(\mathcal{B})}}\boldsymbol{X_{t-k}}\right] \\ &= \frac{\boldsymbol{B(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}}\boldsymbol{X_t} + \boldsymbol{E(\mathcal{B})\varepsilon_{t+k}} + \frac{\boldsymbol{F(\mathcal{B})}}{\boldsymbol{C(\mathcal{B})}}\boldsymbol{Y_t} - \frac{\boldsymbol{F(\mathcal{B})}}{\boldsymbol{A(\mathcal{B})}}\frac{\boldsymbol{B(\mathcal{B})}}{\boldsymbol{C(\mathcal{B})}}\boldsymbol{\mathcal{B}^k X_t} \\ &= \boldsymbol{E(\mathcal{B})\varepsilon_{t+k}} + \frac{\boldsymbol{F(\mathcal{B})}}{\boldsymbol{C(\mathcal{B})}}\boldsymbol{Y_t} + \left[\frac{\boldsymbol{B(\mathcal{B})\left(C(\mathcal{B}) - F(\mathcal{B})\mathcal{B}^k\right)}}{\boldsymbol{A(\mathcal{B})C(\mathcal{B})}}\right]\boldsymbol{X_t}\end{aligned}$$

Da equação (5.11) pode tirar-se ainda que $(\boldsymbol{C(\mathcal{B})} - \boldsymbol{F(\mathcal{B})}\,\boldsymbol{\mathcal{B}\text{^}k)/A(\mathcal{B})} = \boldsymbol{E(\mathcal{B})}$. Introduzindo este resultado no ultimo termo do lado direito da ultima expressão obtém-se

$$\boldsymbol{Y_{t+k}} = \boldsymbol{E(\mathcal{B})\varepsilon_{t+k}} + \frac{\boldsymbol{F(\mathcal{B})}}{\boldsymbol{C(\mathcal{B})}}\boldsymbol{Y_t} + \frac{\boldsymbol{B(\mathcal{B})E(\mathcal{B})}}{\boldsymbol{C(\mathcal{B})}}\boldsymbol{X_t} \quad \textbf{(5.13)}$$

Omitindo o operador de retardo por simplicidade de notação, a expressão anterior fica na forma

$$\boldsymbol{Y_{t+k}} = \boldsymbol{E\varepsilon_{t+k}} + \frac{\boldsymbol{F}}{\boldsymbol{C}}\boldsymbol{Y_t} + \frac{\boldsymbol{BE}}{\boldsymbol{C}}\boldsymbol{X_t}$$

Suponha-se que T_y seja o valor de referência ou alvo, que normalmente é zero, uma vez que Y_t normalmente significa o desvio em relação à referência. Para se obter o controlador de variância mínima, tem que se minimizar o valor esperado dos desvios quadráticos $\left(Y_{t+k} - T_y\right)^2$, ou seja minimizar:

$$\begin{aligned} \boldsymbol{MSE(Y_{t+k})} &= \boldsymbol{E_t\left[\left(Y_{t+k} - T_y\right)^2\right]} \\ &= \underbrace{\boldsymbol{E_t\left[(Y_{t+k} - E_t[Y_{t+k}])^2\right]}}_{\boldsymbol{var(Y_{t+k})}} + \underbrace{\boldsymbol{E_t\left[\left(T_y - E[Y_{t+k}]\right)^2\right]}}_{\boldsymbol{desvio\ estacionário\ quadrático}} \end{aligned}$$

Onde o ultimo termo corresponde ao erro estacionário ou offset da saída do processo. A partir da expressão (5.13) simplificada, e substituindo os desvios futuros por zero (valor esperado de ruído branco) tem-se então que

$$\boldsymbol{E[Y_{t+k}] = \frac{F}{C}Y_t + \frac{BE}{C}X_t}$$

Onde o primeiro termo do lado direito é o valor espectável em t para a variável de saída Y_{t+k} no instante $t+k$ quando o valor da variável de entrada X_t é nulo no instante t, ou seja, se o processo não for controlado. Então, fazendo $T_y = 0$ po, r questões de simplicidade,

$$\begin{aligned} \boldsymbol{MSE(Y_{t+k})} &= \boldsymbol{E_t\left[(Y_{t+k})^2\right] = E_t\left[(Y_{t+k} - E_t[Y_{t+k}])^2\right] + E_t[Y_{t+k}]^2} \\ &= \boldsymbol{E_t\left[(E\varepsilon_{t+k})^2\right] + E_t[Y_{t+k}]^2 \geq \left(1 + e_1^2 + \ \ldots + e_{1-1}^2\right)\sigma_\varepsilon^2} \end{aligned}$$

Onde se obtém a igualdade se se ajustar a variável de entrada X_t de modo a que $E_t[Y_{t+k}] = \mathbf{0}$, ou seja, o valor de X_t tem que ser tal que

$$\boldsymbol{E[Y_{t+k}] = 0 \ \Leftrightarrow \frac{F}{C}Y_t + \frac{BE}{C}X_t = 0}$$

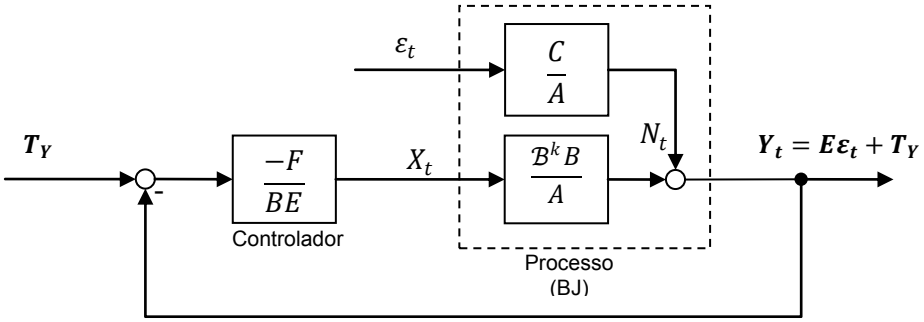
$$\boldsymbol{\Leftrightarrow \quad -\frac{F}{C}Y_t = \frac{BE}{C}X_t}$$

De onde se obtém a lei de controlo de variância mínima ou MMSE:

$$\boldsymbol{X_t = -\frac{F(\mathcal{B})}{B(\mathcal{B})E(\mathcal{B})}Y_t} \qquad \textbf{(5.14)}$$

Substituindo a lei de controlo (5.14) na expressão do modelo do processo (5.13), obtém-se a equação de saída em malha fechada representada pelo diagrama de blocos da:

$$\boldsymbol{Y_{t+k} = E(\mathcal{B})\varepsilon_{t+k} + \frac{F(\mathcal{B})}{C(\mathcal{B})}Y_t + \frac{B(\mathcal{B})E(\mathcal{B})}{C(\mathcal{B})}\left(-\frac{F(\mathcal{B})}{B(\mathcal{B})E(\mathcal{B})}\right)Y_t = E(\mathcal{B})\varepsilon_{t+k}}$$



Da expressão anterior conclui-se que a saída de um sistema constituído por um processo BJ e um controlador MMSE em malha fechada, segue um processo MA(k-1). Isso implica que a saída será estacionária, mas não será ruído branco puro, uma vez que em geral será não correlacionada. Pode ainda concluir-se que, em malha fechada, o desvio médio quadrático será dado por (Del Castillo, 2002):

$$\boldsymbol{MSE(Y_{t+k}) = var(Y_t) = \sigma_\varepsilon^2 \sum_{i=0}^{k-1} e_i^2}$$

5.1.2 Controlador de Variância mínima generalizado

A grande variabilidade a que os factores manipuláveis (entradas) ficam sujeitos sob as leis de controlo de variância mínima, tornam-se geralmente impraticáveis economicamente e/ou mesmo fisicamente. Esse facto deve-se, geralmente, à transferência da variabilidade das saídas (outputs) para as entradas (inputs).

O problema na prática é que uma manipulação excessiva dos factores manipuláveis, na maioria dos casos, tornam-se indesejáveis, não só porque exigem grandes esforços, mas também porque, normalmente, os sistemas são linearizados na vizinhança de um ponto de actuação, e ao afastar-se dele, terá que haver uma adaptação dos próprios parâmetros. Estas limitações podem ser traduzidas custos proporcionais às amplitudes das acções de controlo $|X_t|$. Ou poderão ser impostas restrições na gama de valores que os factores manipuláveis podem tomar; normalmente escalados no intervalo $[-1, 1]$ como é usual em desenho de experiências, que muitas das vezes são utilizados como meio de identificação dos processos. Uma forma indirecta de atingir estes objectivos é restringir a variância dos factores manipuláveis: $var(X_t)$. Pode assim concluir-se que existe a necessidade de um controlador que minimize a variância das variáveis de saída mas sujeito a restrições nas variáveis dos factores manipuláveis. Uma técnica relativamente simples que resolve este problema entre outros foi proposto por Clarke e Gawthrop, pelo que ficou conhecido como o controlador Clarke-Gawthrop.

Os controladores MMSE concentram o seu critério de desempenho na minimização da variância de saída

$$\boldsymbol{J_{MV} = E[Y_t^2]}$$

Que será minimizado seleccionando X_t tal que $\hat{Y}_{t+k|t} = 0$ para todos os instantes t. Uma extensão natural deste critério, e de acordo com o que foi dito no inicio desta secção, será estendê-lo também às variáveis de entrada

$$\boldsymbol{J_I = E\left[Y_{t+k}^2 + \lambda X_t^2\right]}$$

A constante λ pode ser vista como o multiplicador de Lagrange de $E[X_t^2] \leq c$ que é a restrição do *MSE* do factor manipulável. Se X_t está escalado de forma centrada em zero, e se o sistema no estado livre (não controlado) não flutua, então $MSE[X_t] = var(X_t)$.

Para minimizar J_I, Clarke e Gawthrop propôs a decomposição de Y_{t+k} em

$$\boldsymbol{Y_{t+k} = \hat{Y}_{t+k|t} + \tilde{Y}_{t+k|t}}$$

Em que o ultimo termo é o erro de previsão, que é independente de X_t, e que portanto é apenas função de erros futuros em relação ao instante t. Introduzindo esta ultima expressão, J_I será dado por (Del Castillo, 2002):

$$\boldsymbol{E\left[Y_{t+k}^2 + \lambda X_t^2\right] = E\left[\hat{Y}_{t+k|t}^2 + 2\hat{Y}_{t+k|t}\tilde{Y}_{t+k|t} + \tilde{Y}_{t+k|t}^2 + \lambda X_t^2\right]}$$

Como os termos cruzados são nulos e $E\left[\tilde{Y}_{t+k|t}^2\right] = \boldsymbol{var}\left(\tilde{Y}_{t+k|t}^2\right)$, a expressão anterior reduz-se a

$$\boldsymbol{E\left[Y_{t+k}^2 + \lambda X_t^2\right] = var\left(\tilde{Y}_{t+k|t}^2\right) + E\left[\hat{Y}_{t+k|t}^2 + \lambda X_t^2\right]}$$

Como $\boldsymbol{var\left(\tilde{Y}_{t+k|t}^2\right)}$ não é função de X_t, para se minimizar J_I torna-se suficiente minimizar apenas o ultimo termo da expressão anterior

$$\boldsymbol{J_2 = E\left[\hat{Y}_{t+k|t}^2 + \lambda X_t^2\right] = \int_{-\infty}^{\infty} \left(\hat{Y}_{t+k|t}^2 + \lambda X_t^2\right) f(X_t)\, d\, X_t}$$

Onde $\boldsymbol{f(X_t)}$ é a função densidade probabilidade de $\boldsymbol{X_t}$. Assim esta expressão reconhece o facto de que o factor de controlo é uma variável aleatória que depende de $\boldsymbol{Y_t}$ que por seu lado depende dos erros $\boldsymbol{\varepsilon_t}$. Como se pode verificar, surge uma dificuldade: A densidade de $\boldsymbol{X_t}$ depende da lei de controlo por realimentação (*feedback*) que se pretende determinar em primeiro lugar.

Por outro lado, Clarke e Gawthrop propôs minimizar o valor esperado condicional:

$$\boldsymbol{J_3 = E\left[\hat{Y}_{t+k}^2 + \lambda X_t^2 | Y_t, Y_{t-1}, \ldots\right] = E\left[Y_{t+k}^2 + \lambda X_t^2\right] = \hat{Y}_{t+k|t}^2 + \lambda X_t^2} \quad \textbf{(5.15)}$$

Onde a ultima igualdade surge de se ter tomado o valor esperado condicionada às observações presentes e passadas. Se as observações $\{Y_i\}_{i \leq t}$ são conhecidas, então $\boldsymbol{X_t}$ deixa de ser uma variável aleatória, pelo que e o valor esperado condicional de $\boldsymbol{Y_{t+k}^2}$ pode ser então determinado usando o critério da mínima media dos erros quadráticos MMSE (5.13):

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{t+k|t}} = \boldsymbol{E_t}[\boldsymbol{Y_{t+k}}] = \frac{\boldsymbol{F}(\boldsymbol{\mathcal{B}})}{\boldsymbol{C}(\boldsymbol{\mathcal{B}})}\boldsymbol{Y_t} + \frac{\boldsymbol{B}(\boldsymbol{\mathcal{B}})\boldsymbol{E}(\boldsymbol{\mathcal{B}})}{\boldsymbol{C}(\boldsymbol{\mathcal{B}})}\boldsymbol{X_t}$$

Inserindo este resultado em (5.15) obtém-se

$$\boldsymbol{J_3} = \left[\frac{\boldsymbol{F}}{\boldsymbol{C}}\boldsymbol{Y_t} + \frac{\boldsymbol{BE}}{\boldsymbol{C}}\boldsymbol{X_t}\right]^2 + \boldsymbol{\lambda X_t^2}$$

onde, por uma questão de simplicidade, se omitiu o operador de atraso. Para minimizar $\boldsymbol{J_3}$ procura-se o mínimo da função recorrendo aos zeros da derivada:

$$\frac{\boldsymbol{dJ_3}}{\boldsymbol{dX_t}} = \frac{\boldsymbol{d}}{\boldsymbol{dX_t}}\left(\widehat{\boldsymbol{Y}}^{\boldsymbol{2}}_{\boldsymbol{t+k|t}} + \boldsymbol{\lambda X_t^2}\right) = \boldsymbol{2}\frac{\boldsymbol{d}\left(\widehat{\boldsymbol{Y}}_{\boldsymbol{t+k|t}}\right)}{\boldsymbol{dX_t}}\widehat{\boldsymbol{Y}}_{\boldsymbol{t+k|t}} + \boldsymbol{2\lambda X_t}$$

Como os únicos coeficientes dos polinómios *B, F* e *C* que estão envolvidos no termo derivativo $d\left(\hat{Y}_{t+k|t}\right)/dX_t$ são os coeficientes de ordem zero, tem-se

$$\frac{\boldsymbol{d}\left(\widehat{\boldsymbol{Y}}_{\boldsymbol{t+k|t}}\right)}{\boldsymbol{dX_t}} = \frac{\boldsymbol{B}(\boldsymbol{0})\boldsymbol{E}(\boldsymbol{0})}{\boldsymbol{C}(\boldsymbol{0})} = \boldsymbol{b_0}$$

Deste modo, tem-se finalmente que

$$\frac{\boldsymbol{dJ_3}}{\boldsymbol{dX_t}} = \boldsymbol{0} \;\Rightarrow\; \boldsymbol{2b_0}\left(\frac{\boldsymbol{F}}{\boldsymbol{C}}\boldsymbol{Y_t} + \frac{\boldsymbol{BE}}{\boldsymbol{C}}\boldsymbol{X_t}\right) + \boldsymbol{2\lambda X_t} = \boldsymbol{0}$$

Resolvendo em ordem a $\boldsymbol{X_t}$ obtém-se finalmente o controlador de variância mínima generalizada de Clarke-Gawthrop:

$$\boldsymbol{X_t} = \frac{\boldsymbol{b_0 F}(\boldsymbol{\mathcal{B}})}{\boldsymbol{b_0 B}(\boldsymbol{\mathcal{B}})\boldsymbol{E}(\boldsymbol{\mathcal{B}}) + \boldsymbol{\lambda C}(\boldsymbol{\mathcal{B}})}\boldsymbol{Y_t} \qquad \textbf{(5.16)}$$

6 Análise Multivariada

6.1 Séries temporais Multivariadas

A análise multivariada de séries temporais é o estudo de modelos estatísticos e métodos de análise que descrevem a relação entre várias séries temporais. Ou seja passa-se de uma variável z_t usada para quantificar a saída de um determinado processo num determinado instante t no caso univariado para uma variável vectorial $Z_t = (z_{1t},\ z_{2t}, \dots\ z_{kt})$ que quantifica as diversas variáveis mensuráveis envolvidas no processo. Os processos multi-variáveis têm interesse nas mais variadas áreas tais com engenharia, física, ciências da Terra tais como meteorologia ou geofísica, economia e finanças. Por exemplo, em engenharia, pode-se estar interessado em estudar comportamentos simultâneos de corrente e tensão, temperatura, pressão e volume, ao passo que em economia pode-se estar interessado simultaneamente em taxas de juro, disponibilidade financeira, desemprego etc.

No estudo de processos multi-variados, torna-se necessário a existência de uma estrutura que descreva não apenas as propriedades das séries individuais mas também a possibilidade de relações cruzadas entre as séries. Estas relações são frequentemente estudadas e detectadas da consideração das estruturas de correlação entre as várias componentes da série. Os objectivos da análise e modelação das séries em conjunto são compreender a relação dinâmica ao longo do tempo entre as séries e melhorar a precisão da previsão para as séries individuais recorrendo à informação adicional proveniente da relação entre as séries.

A maioria fundamental dos conceitos teóricos descritos nas secções anteriores para as séries temporais univariadas estendem-se aos sistemas multivariados, mas surgirão novos problemas e desafios na modelação e análise das séries temporais multivariadas devido à grande complexidade dos modelos e parametrizações na situação vectorial.

6.1.1 Series temporais multivariadas estacionárias

Seja então $\boldsymbol{Z_t} = (Z_{1t},\ \dots\ , Z_{kt}),\ t = 0, \pm1, \pm2, \dots$ um vector série temporal de dimensão k, a escolha de uma determinada serie componente Z_{it} incluida em $\boldsymbol{Z_t}$, dependerá do assunto e compreensão do sistema em estudo, mas está implícito que a serie componente estará inter-relacionada tanto contemporaneamente, ao longo das séries componentes, com ao longo do tempo (das *lags* de tempo). O principal interesse da análise das series temporais multivariadas será a representação e modelação das relações dinâmicas entre as diversas series componentes. Tal como no caso univariado, um conceito importante na análise, e representação dos modelos de séries temporais é a estacionaridade.

O processo $\{\boldsymbol{Z_t}\}$ é estacionário se a distribuição de probabilidade dos vectores $\left(Z_{t_1}, Z_{t_2}, \dots\ , Z_{t_m}\right)$ e $\left(Z_{t_1+l}, Z_{t_2+l},\ \dots\ , Z_{t_m+l}\right)$ é a mesma para tempos arbitrários $t_1, t_2, \dots\ t_m$, qualquer que seja m ou l, ou seja, a distribuição de probabilidade das observações de um processo vectorial estacionário é invariante no tempo. Consequentemente, assumindo-se que os primeiro e segundo momentos existem, para um processo estacionário ter-se-á $E[\boldsymbol{Z_t}] = \boldsymbol{\mu}$ constante, qualquer que seja t, em que $\boldsymbol{\mu} = (\boldsymbol{\mu_1}, \boldsymbol{\mu_2},\ \dots\ , \boldsymbol{\mu_k})'$ é o vector da média do

processo. Adicionalmente, o vector $\boldsymbol{Z_t}$ deverá ter uma matriz de covariância constante definida por $\Sigma_z = \Gamma_0 = E[(Z_t - \mu)(Z_t - \mu)^{'}]$. Ver (3.2.2-Processo estacionário).

6.1.1.1 Covariância e correlação de um processo vectorial estacionário

Para um processo estacionário $\{\boldsymbol{Z_t}\}$, a covariância entre $z_{i,t}$ e $z_{j,t+l}$ tem de depender apenas de da *lag* l e não do tempo t.

Para os componentes $i = 1, \dots, k,\ j = 1, \dots, k$, de um processo estacionário vectorial $\boldsymbol{Z_t}$ de dimensão k, e para as *lags* $j = 0, \pm1, \pm2, \dots$, a covariância cruzada entre o componente i e o componente j na lag l é definida pela expressão:

$$\boldsymbol{\gamma_{ij}(l) = Cov(Z_{it}, Z_{j,t+l}) = E[(z_{i,t} - \mu_i)(z_{j,t+l} - \mu_j)]}$$

e a matriz de dimensão kxk de covariância cruzada da *lag* l é definida por

$$\boldsymbol{\Gamma(l) = [(Z_t - \mu)(Z_t - \mu)^{'}] = \begin{bmatrix} \gamma_{11}(l) & \gamma_{12}(l) & \cdots & \gamma_{1k}(l) \\ \gamma_{21}(l) & \gamma_{22}(l) & \cdots & \gamma_{2k}(l) \\ \vdots & \vdots & \cdots & \vdots \\ \gamma_{k1}(l) & \gamma_{k2}(l) & \cdots & \gamma_{kk}(l) \end{bmatrix}} \quad \textbf{(6.1)}$$

para $l = 0, \pm1, \pm2, \dots$. As correspondentes correlações cruzadas da *lag* l são definidas por

$$\boldsymbol{\rho_{ij}(l) = Corr(Z_{it}, Z_{j,t+l}) = \frac{\gamma_{ij}(l)}{[\gamma_{ii}(0)\gamma_{jj}(0)]^{1/2}}}$$

Com $\gamma_{ii}(0) = Var(Z_{it})$ obviamente. Da mesma forma, a matriz de correlação cruzada da *lag* l é definida por

$$\boldsymbol{\rho(l) = V^{-1/2}\Gamma(l)V^{-1/2} = \begin{bmatrix} \rho_{11}(l) & \rho_{12}(l) & \cdots & \rho_{1k}(l) \\ \rho_{21}(l) & \rho_{22}(l) & \cdots & \rho_{2k}(l) \\ \vdots & \vdots & \cdots & \vdots \\ \rho_{k1}(l) & \rho_{k2}(l) & \cdots & \rho_{kk}(l) \end{bmatrix}} \quad \textbf{(6.2)}$$

para $l = 0, \pm1, \pm2, \dots$, onde $V^{-1/2} = \boldsymbol{Diag}\{\gamma_{11}(0)^{-1/2}, \dots, \gamma_{kk}(0)^{-1/2}\}$. Deste modo, para $i = j$, $\rho_{ii}(l) = \rho_{ii}(-l)$ é a função autocorrelação da *i*-ésima série Z_{it}, e para $i \neq j$, $\rho_{ij}(l) = \rho_{ji}(-l)$ é a função de correlação cruzada entre as séries Z_{it} e Z_{jt}. De notar que $\Gamma(l)^{'} = \Gamma(-l)$ e $\rho(l)^{'} = \rho(-l)$, uma vez que $\gamma_{ij}(l) = \gamma_{ji}(-l)$ (Reinsel, 1997), ver Figura 4-11 para o caso bivariado. Adicionalmente, as matrizes de correlação cruzada $\Gamma(l)$ e de covariância cruzada $\rho(l)$ tem a propriedade de serem não negativas definidas, uma vez que

$$\boldsymbol{var\left[\sum_{i=1}^{n} b_i^{'} Z_{t-i}\right] = \sum_{i=1}^{n}\sum_{i=1}^{n} b_i^{'}\Gamma(i-j)b_j \geq 0}$$

Para qualquer inteiro positivo n e qualquer vector constante de dimensão k $b_1, \dots, b_n$ (Reinsel, 1997)

6.1.1.2 Desvios Normais Multivariados

Processo ruído branco vectorial

O exemplo base de um processo estacionário vectorial ou multivariado, é o processo ruído branco multivariado, que funciona como unidade base estrutural para processos vectoriais (multivariados).

O processo vectorial ruído branco é definido como uma sequência de vectores aleatórios independentes, $\dots, a_1, \dots, a_t, \dots$ em que $a_t = (a_{1t}, \dots, a_{kt})'$, tal que $E[a_t] = 0$, $E[a_t a_t'] = \Sigma$, onde Σ é a matriz de covariância kxk assumida positiva definida, e $E[a_t a_{t+l}'] = 0$ para $l \neq 0$ devido à independência. Consequentemente as suas matrizes de covariância são dadas por

$$\boldsymbol{\Gamma(l) = E[a_t a_{t+l}'] = \begin{cases} \Sigma & se\ l = 0 \\ 0 & se\ l \neq 0 \end{cases}}$$

De notar que, apesar dos vectores ruído branco serem independentes e identicamente distribuídos (IID) entre eles, os k elementos componentes de cada vector não gozam da mesma propriedade.

Desvios normais multivariados

Um desvio aleatório vectorial de dimensão k é um ponto num espaço de dimensão k. As suas coordenadas são um vector em que cada uma das suas componentes é aleatória, mas que, em geral, não são independentes ou identicamente distribuídas. O caso especial de desvios normais multivariados é definido função densidade probabilidade Gaussiana multidimensional

$$\boldsymbol{N(x|\mu,\Sigma) = \frac{1}{(2\pi)^{k/2} det(\Sigma)^{1/2}} exp\left[-\frac{1}{2}(x-\mu)' \cdot \Sigma^{-1} \cdot (x-\mu)\right]}$$

Onde o parâmetro $\boldsymbol{\mu}$ é um vector que, como normalmente, é a média da distribuição, e o parâmetro $\boldsymbol{\Sigma}$ é uma matriz simétrica, positiva definida, que, como normalmente, é a covariância da distribuição.

Existe um modo bastante geral para construir um vector de desvios $\boldsymbol{x}$ com parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, partindo de um vector $\boldsymbol{y}$ de desvios aleatórios independentes de média zero e variância unitária: primeiro utilizar a decomposição de Cholesky para factorizar $\boldsymbol{\Sigma}$ numa matriz triangular L multiplicada pela sua transposta

$$\boldsymbol{\Sigma = LL^T} \quad \textbf{(6.3)}$$

Este passo é sempre possível porque $\boldsymbol{\Sigma}$ é uma matriz positiva definida, e apenas se necessita de executar este passo para cada matriz $\boldsymbol{\Sigma}$ distinta. A seguir, sempre que se necessitar um novo desvio $\boldsymbol{x}$, construir um vector $\boldsymbol{y}$ com desvios normais independentes de média zero e variância unitária, e então construir (Teukolsky, Vetterling, & Flannery, 2007)

$$\boldsymbol{x} = \boldsymbol{L}\boldsymbol{y} + \boldsymbol{\mu} \tag{6.4}$$

Retirar a componente de correlação de variáveis aleatórias multivariadas

De modo similar, mas em sentido contrário, dada a decomposição (6.3) e um vector $\boldsymbol{x}$, constituído por ruído branco multivariado, cujas componentes tem covariância $\boldsymbol{\Sigma}$ e média $\boldsymbol{\mu}$ conhecidas, tem-se que o vector dado por

$$\boldsymbol{y} = \boldsymbol{L}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \tag{6.5}$$

tem componentes não correlacionadas com variância unitária.

6.1.1.3 Filtros lineares

Um filtro linear (invariante no tempo) que relacione uma série de entrada $\boldsymbol{X_t}$, de dimensão r, com uma serie de saída $\boldsymbol{Z_t}$, de dimensão k é dado pela forma

$$\boldsymbol{Z_t} = \sum_{\mathbf{j}=-\infty}^{\infty} \boldsymbol{\Psi_j}\, \boldsymbol{X_{t-j}}$$

onde $\boldsymbol{\Psi_j}$ é uma matriz $\boldsymbol{k} \times \boldsymbol{r}$. O filtro é fisicamente realizável ou causal quando $\boldsymbol{\Psi_j} = \mathbf{0}$ para $\boldsymbol{j} < 0$, tal que $\boldsymbol{Y_t} = \sum_{\mathbf{j}=-\infty}^{\infty} \boldsymbol{\Psi_j}\, \boldsymbol{X_{t-j}}$ se pode expressar apenas em termos de valores passados e presente do processo de entrada $\{\boldsymbol{X_t}\}$. O filtro diz-se estável se $\sum_{\mathbf{j}=-\infty}^{\infty} \|\boldsymbol{\Psi_j}\| < \infty$ onde $\|A\|$ significa a norma da matriz A definida por $\|A\|^2 = tr\{A'A\}$. Quando o filtro é estável e a serie de entrada $\boldsymbol{X_t}$ é estacionária com matriz de covariância cruzada $\Gamma_{xx}(l)$, a saída $\boldsymbol{Z_t}$ é um processo estacionário. A matriz de covariância cruzada do processo estacionário $\{\boldsymbol{Z_t}\}$ é então dada por

$$\boldsymbol{\Gamma_{zz}(l)} = \boldsymbol{cov}[\boldsymbol{Z_t}, \boldsymbol{Z_{t+l}}] = \sum_{\boldsymbol{i}=-\infty}^{\infty} \sum_{\boldsymbol{j}=-\infty}^{\infty} \boldsymbol{\Psi_j} \Gamma_{xx}(l + i - j)\, \boldsymbol{\Psi'_j}$$

6.1.2 Representação de modelos lineares para processos vectoriais estacionários

Um processo vectorial $\{\boldsymbol{Z_t}\}$ pode ser representado como um processo média móvel (MA) infinito

$$\boldsymbol{Z_t} = \boldsymbol{\mu} + \sum_{\mathbf{j}=\mathbf{0}}^{\infty} \boldsymbol{\Psi_j}\, \boldsymbol{\varepsilon_{t-j}} = \boldsymbol{\mu} + \boldsymbol{\Psi}(\boldsymbol{\mathcal{B}})\boldsymbol{\varepsilon_t}, \qquad \boldsymbol{\Psi_0} = \mathbf{I} \tag{6.6}$$

onde $\boldsymbol{\Psi(\mathcal{B})} = \sum_{\mathbf{j=0}}^{\infty} \boldsymbol{\Psi_j \mathcal{B}^j}$ é uma matriz $\boldsymbol{k \times k}$ do operador de atraso, e os coeficientes da matriz $\boldsymbol{\Psi_j}$ satisfazem a condição $\sum_{\mathbf{j=0}}^{\infty} \|\boldsymbol{\Psi_j}\| < \infty$**.** Na expressão (6.6), $\boldsymbol{\varepsilon_t} = (\boldsymbol{\varepsilon_{1t}}, \dots, \boldsymbol{\varepsilon_{kt}})$ forma o vector ruído branco.

6.1.2.1 Representações de modelos VARMA (Vector Autoregressive-Moving Average)

Suponha-se que a matriz $\boldsymbol{\Psi(\mathcal{B})}$ pode ser representada (pelo menos de forma aproximada) pelo produto de duas matrizes na forma $\boldsymbol{\Phi^{-1}(\mathcal{B})\Theta(\mathcal{B})}$, onde $\boldsymbol{\Phi(\mathcal{B})} = \boldsymbol{I - \Phi_1 \mathcal{B} - \Phi_2 \mathcal{B}^2 - \cdots - \Phi_p \mathcal{B}^p}$ e $\boldsymbol{\Theta(\mathcal{B})} = \boldsymbol{I - \Theta_1 \mathcal{B} - \Theta_2 \mathcal{B}^2 - \cdots - \Theta_q \mathcal{B}^q}$ são polinómios em $\boldsymbol{\mathcal{B}}$ de matrizes de ordem finita, e $\boldsymbol{\Phi_i}$ e $\boldsymbol{\Theta_i}$ são matrizes $\boldsymbol{k \times k}$. Pode-se então considerar uma classe de modelos lineares para series temporais vectoriais $\boldsymbol{Z_t}$ definidas pela relação $\boldsymbol{\Phi(\mathcal{B})(Z_t - \mu)} = \boldsymbol{\Theta(\mathcal{B})\varepsilon_t}$, ou seja

$$\boldsymbol{(Z_t - \mu) - \sum_{j=1}^{p} \Phi_j (Z_{t-j} - \mu) = \varepsilon_t - \sum_{j=1}^{q} \Theta_j \varepsilon_{t-j}} \qquad \textbf{(6.7)}$$

onde $\boldsymbol{\varepsilon_t}$ é um vector ruido branco com média zero e matriz de covariância $\Sigma = E[\varepsilon_t \varepsilon_t^{'}]$. Este modelo é um generalização natural dos modelos univariados ARMA(p,q) que será de grande utilidade na modelação de série temporais vectoriais.

Um processo vectorial $\{\boldsymbol{Z_t}\}$ estacionário ARMA(p,q) definido pela relação (6.7) diz-se causal se puder ser representado na forma (6.6) para qualquer $t = 0, \pm 1, \pm 2, \dots$, com $\sum_{\mathbf{j=0}}^{\infty} \|\boldsymbol{\Psi_j}\| < \infty$. Um processo vectorial ARMA(p,q) diz-se invertível se puder ser representado na forma

$$\boldsymbol{(Z_t - \mu) - \sum_{j=1}^{\infty} \Pi_j (Z_{t-j} - \mu) = \varepsilon_t} \qquad \textbf{(6.8)}$$

ou $\boldsymbol{\Pi(\mathcal{B})(Z_t - \mu) = \varepsilon_t}$ onde $\boldsymbol{\Pi(\mathcal{B})} = \boldsymbol{I - \sum_{j=1}^{\infty} \Pi_j \mathcal{B}^j}$ com $\sum_{\mathbf{j=0}}^{\infty} \|\boldsymbol{\Pi_j}\| < \infty$.

Estacionaridade

Um processo vectorial $\{\boldsymbol{Z_t}\}$ descrito por um modelo ARMA(*p,q*) na forma (6.7), é estacionário se todas as raízes de $\boldsymbol{det[\Phi(\mathcal{B}) = 0]}$ são maiores que um em valor absoluto. Nesse caso, o processo $\{\boldsymbol{Z_t}\}$, pode ser representado na forma de média móvel (MA) infinita como em (6.6), com $\boldsymbol{\Psi(\mathcal{B})} = \boldsymbol{\Phi^{-1}(\mathcal{B})\Theta(\mathcal{B})}$ representando a serie matricial convergente para $\boldsymbol{|\mathcal{B}| \leq 1}$.

Invertibilidade

De forma similar, um processo vectorial $\{\boldsymbol{Z_t}\}$ descrito por um modelo ARMA(p,q) na forma (6.7), diz-se invertível se todas as raízes de $\boldsymbol{det[\Theta(\mathcal{B}) = 0]}$ são maiores que um em valor absoluto (Reinsel, 1997). Nesse caso, o processo é invertível com

$$\boldsymbol{\Pi(\mathcal{B}) = \Theta^{-1}(\mathcal{B})\Phi(\mathcal{B}) = I - \sum_{j=1}^{\infty} \Pi_j \mathcal{B}^j}$$

Representando a série matricial convergente para $\boldsymbol{|\mathcal{B}| \leq 1}$ e $\{\boldsymbol{Z_t}\}$ possui uma representação autoregressiva infinita como em (6.8).

Segue-se então que para um processo estacionário ARMA(p,q), os coeficientes matriciais $\boldsymbol{\Psi}(\boldsymbol{\mathcal{B}})$ da representação MA infinita (6.6) são determinados a partir da relação $\boldsymbol{\Phi}(\boldsymbol{\mathcal{B}})\boldsymbol{\Psi}(\boldsymbol{\mathcal{B}}) = \boldsymbol{\Theta}(\boldsymbol{\mathcal{B}})$, e consequentemente, igualando os coeficientes com o mesmo expoente, satisfazem a relação

$$\boldsymbol{\Psi_j} = \boldsymbol{\Phi_1}\boldsymbol{\Psi_{j-1}} + \boldsymbol{\Phi_2}\boldsymbol{\Psi_{j-2}} + \cdots + \boldsymbol{\Phi_p}\boldsymbol{\Psi_{j-p}} - \boldsymbol{\Theta_j} \qquad \boldsymbol{j = 1, 2, \ldots}$$

onde $\boldsymbol{\Psi_0} = \boldsymbol{I}$, $\boldsymbol{\Psi_j} = \boldsymbol{0}$ para $j < 0$, e $\boldsymbol{\Theta_j} = \boldsymbol{0}$ para $j > q$.

Reciprocamente, nas condições de invertibilidade, os coeficientes de ponderação $\boldsymbol{\Pi_j}$ na representação infinita AR (6.8) são determinados a partir da relação $\boldsymbol{\Theta}(\boldsymbol{\mathcal{B}})\boldsymbol{\Pi}(\boldsymbol{\mathcal{B}}) = \boldsymbol{\Phi}(\boldsymbol{\mathcal{B}})$, e consequentemente, igualando os coeficientes com o mesmo expoente, satisfazem a relação

$$\boldsymbol{\Pi_j} = \boldsymbol{\Theta_1}\boldsymbol{\Pi_{j-1}} + \boldsymbol{\Theta_2}\boldsymbol{\Pi_{j-2}} + \cdots + \boldsymbol{\Theta_q}\boldsymbol{\Pi_{j-q}} - \boldsymbol{\Phi_j} \qquad \boldsymbol{j = 1, 2, \ldots}$$

onde $\boldsymbol{\Pi_0} = -\boldsymbol{I}$, $\boldsymbol{\Pi_j} = \boldsymbol{0}$ para $j < 0$, e $\boldsymbol{\Phi_j} = \boldsymbol{0}$ para $j > p$.

Matrizes de covariância de processos vectoriais ARMA

Um processo estacionário ARMA(*p,q*) dado por (6.7) pode ter uma representação MA infinita dada por (6.6). Então, a partir de (6.6), é fácil de verificar que (Box, Jenkins, & Reinsel, 2008)

$$\boldsymbol{E}\left[\boldsymbol{Z_{t-l}}\boldsymbol{\varepsilon_{t-j}'}\right] = \begin{cases} \boldsymbol{0} & \boldsymbol{se\ j < l} \\ \boldsymbol{\Psi_{j-l}\Sigma} & \boldsymbol{se\ j \geq l} \end{cases}$$

Deste modo, torna-se fácil determinar, a partir de (6.7) que as matrizes de covariância $\Gamma(l) = E[(\boldsymbol{Z_{t-l}} - \boldsymbol{\mu})(\boldsymbol{Z_t} - \boldsymbol{\mu})']$ do processo $\{\boldsymbol{Z_t}\}$ satisfazem a relação

$$\boldsymbol{Cov}(\boldsymbol{Z_{t-l}}, \boldsymbol{Z_t}) = \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{Cov}\left(\boldsymbol{Z_{t-l}}, \boldsymbol{Z_{t-j}}\right)\boldsymbol{\Phi_j'} + \boldsymbol{Cov}(\boldsymbol{Z_{t-l}}, \boldsymbol{\varepsilon_t}) - \sum_{\boldsymbol{j=1}}^{\boldsymbol{q}} \boldsymbol{Cov}(\boldsymbol{Z_{t-l}}, \boldsymbol{\varepsilon_t})\,\boldsymbol{\Theta_j'}$$

e, consequentemente,

$$\boldsymbol{\Gamma(l)} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\Gamma(l-j)}\,\boldsymbol{\Phi_j'} - \sum_{\boldsymbol{j=l}}^{\boldsymbol{q}} \boldsymbol{\Psi_{j-l}}\,\boldsymbol{\Sigma}\,\boldsymbol{\Theta_j'} \qquad \boldsymbol{j = 0, 1, \ldots, q} \qquad \textbf{(6.9)}$$

Com a convenção de que $\boldsymbol{\Theta_0} = -\boldsymbol{I}$, e que $\boldsymbol{\Gamma(l)} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\Gamma(l-j)}\,\boldsymbol{\Phi_j'}$ para $l > q$. Deste modo, as matrizes de covariância $\boldsymbol{\Gamma(l)}$, podem ser desenvolvidas em termos dos parâmetros matriciais AR e MA, $\boldsymbol{\Phi_j}$ e $\boldsymbol{\Theta_j}$, e $\boldsymbol{\Sigma}$, através desta forma recursiva.

Modelo vectorial autoregressivo

Se num modelo vectorial ARMA(p,q), a ordem da componente média móvel for zero então ter-se-á um modelo vectorial autoregressivo puro vector ARMA(p,0) ou vector AR(p) ou ainda VAR(q) dado pela expressão

$$(\boldsymbol{Z_t} - \boldsymbol{\mu}) - \sum_{j=1}^{p} \boldsymbol{\Phi_j}(\boldsymbol{Z_{t-j}} - \boldsymbol{\mu}) = \boldsymbol{\varepsilon_t}$$

Tal como nos processos univariados, as condições de estacionaridade de um processo vectorial AR(p) são as mesmas que para um processo ARMA(*p,q*), ou seja $\{\boldsymbol{Z_t}\}$ é estacionário se todas as raízes de $\boldsymbol{det}[\boldsymbol{\Phi}(\boldsymbol{\mathcal{B}}) = \boldsymbol{0}]$ são maiores que um em valor absoluto. Nesse caso, o processo $\{\boldsymbol{Z_t}\}$ tem uma representação MA infinita $\boldsymbol{Z_t} - \boldsymbol{\mu} = \boldsymbol{\Phi}(\boldsymbol{\mathcal{B}})^{-1}\boldsymbol{\varepsilon_t} = \sum_{j=0}^{\infty} \boldsymbol{\Psi_j}\, \boldsymbol{\varepsilon_{t-j}}$, em que $\boldsymbol{\Psi}(\boldsymbol{\mathcal{B}}) = \boldsymbol{\Phi}(\boldsymbol{\mathcal{B}})^{-1}$.

Para um modelo vectorial autoregressivo puro VAR(*p*), a expressão das matrizes de covariância (6.9) reduz-se à matriz das equações de *Yule-Walker* dadas por

$$\boldsymbol{\Gamma}(\boldsymbol{l}) = \sum_{j=1}^{p} \boldsymbol{\Gamma}(\boldsymbol{l}-\boldsymbol{j})\, \boldsymbol{\Phi_j'} \qquad \boldsymbol{para\ l} = \boldsymbol{1, 2, \ldots} \tag{6.10}$$

Com $\boldsymbol{\Gamma}(\boldsymbol{0}) = \sum_{j=1}^{p} \boldsymbol{\Gamma}(\boldsymbol{l}-\boldsymbol{j})\, \boldsymbol{\Phi_j'} + \boldsymbol{\Sigma}$. Estas equações podem ser de grande utilidade para determinar $\boldsymbol{\Gamma}(\boldsymbol{l}),\ \boldsymbol{l} = \boldsymbol{0, 1, 2, \ldots\ , p,}$ em termos dos parâmetros matriciais AR, $\boldsymbol{\Phi_j}$ e $\boldsymbol{\Sigma}$.

Reciprocamente, conhecida ou estimada a ordem p, as matrizes $\boldsymbol{\Phi_1, \ldots\ , \Phi_p}$ e $\boldsymbol{\Sigma}$ podem ser determinadas a partir de $\boldsymbol{\Gamma(0), \Gamma(1),\ \ldots\ , \Gamma(p)}$ resolvendo o sistema de equações matriciais de *Yule-Walker*

$$\sum_{j=1}^{p} \boldsymbol{\Gamma}(\boldsymbol{l}-\boldsymbol{j})\, \boldsymbol{\Phi_j'} = \boldsymbol{\Gamma}(\boldsymbol{l}) \qquad \boldsymbol{para\ l} = \boldsymbol{1, 2, \ldots\ , p} \tag{6.11}$$

Estas equações podem serem escritas na forma matricial como $\boldsymbol{\Gamma_p \Phi} = \boldsymbol{\Gamma_{(p)}}$ com solução

$$\boldsymbol{\Phi} = \boldsymbol{\Gamma_p^{-1} \Gamma_{(p)}} \tag{6.12}$$

onde $\boldsymbol{\Phi} = (\boldsymbol{\Phi_1},\ \boldsymbol{\Phi_1},\ \ldots\ , \boldsymbol{\Phi_p})'$ é uma matriz $kp \times k$, $\boldsymbol{\Gamma_{(p)}} = (\boldsymbol{\Gamma}(\boldsymbol{1})',\ \boldsymbol{\Gamma}(\boldsymbol{2})',\ \ldots\ , \boldsymbol{\Gamma}(\boldsymbol{p})')'$ é uma matriz $kp \times k$ e $\boldsymbol{\Gamma_p}$ é uma matriz bloco $kp \times kp$ em que a matriz elementar $p \times k$ índice (i, j) é composta pela matriz de covariância $\boldsymbol{\Gamma}(\boldsymbol{i}-\boldsymbol{j})$.

Uma vez determinado a matriz coeficiente $\boldsymbol{\Phi_j}$ a partir de (6.11), a matriz de covariância $\boldsymbol{\Sigma}$ pode ser determinada a partir da expressão (Reinsel, 1997)

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}(\boldsymbol{0}) - \sum_{j=1}^{p} \boldsymbol{\Gamma}(-\boldsymbol{j})\, \boldsymbol{\Phi_j'} = \boldsymbol{\Gamma}(\boldsymbol{0}) - \boldsymbol{\Gamma_{(p)}'\Phi} = \boldsymbol{\Gamma}(\boldsymbol{0}) - \boldsymbol{\Gamma_{(p)}'\Gamma_p^{-1}\Gamma_{(p)}} = \boldsymbol{\Gamma}(\boldsymbol{0}) - \boldsymbol{\Phi'\Gamma_p\Phi}$$

Modelo vectorial média móvel

Se num modelo vectorial ARMA(*p,q*), a ordem da componente autoregressiva for zero então ter-se-á um modelo média móvel puro vector ARMA(0,q) ou vector MA(q) ou ainda VMA(q) dado pela expressão

$$\boldsymbol{Z_t = \mu + \Theta(\mathcal{B})\varepsilon_t = \mu + \varepsilon_t - \sum_{j=1}^{q} \Theta_j\, \varepsilon_{t-j}}$$

Tal como nos processos univariados, as condições de invertibilidade de um processo vectorial MA(*q*) são as mesmas que para um processo ARMA(*p,q*), ou seja $\{\boldsymbol{Z_t}\}$ é invertível se todas as raízes de $\boldsymbol{det[\Theta(\mathcal{B}) = 0]}$ são maiores que um em valor absoluto. Nesse caso, o processo $\{\boldsymbol{Z_t}\}$ tem uma representação AR infinita $\boldsymbol{\Theta(\mathcal{B})^{-1}(Z_t - \mu) = \varepsilon_t}$ ou $\boldsymbol{(Z_t - \mu) - \sum_{j=1}^{\infty} \Pi_j(Z_{t-j} - \mu) = \varepsilon_t}$, com $\boldsymbol{\Pi(\mathcal{B}) = \Theta(\mathcal{B})^{-1}}$.

Para um modelo vectorial média novel puro VMA(*q*), a expressão das matrizes de covariância (6.9) reduz-se a

$$\boldsymbol{\Gamma(l) = \sum_{h=0}^{q-l} \Theta_h\, \Sigma\, \Theta_{h+l}^{'}} \qquad \textbf{(6.13)}$$

Para $l = 0, 1, \ldots, q$, com $\boldsymbol{\Theta_0 = -I}$, $\boldsymbol{\Gamma(l) = \Gamma(-l)'}$ e $\boldsymbol{\Gamma(l) = 0}$ para $l > q$ (Reinsel, 1997). Verifica-se assim que para um processo MQ(*q*), todas as correlações cruzadas são nulas para *lags* maiores que *q*. Reciprocamente, a expressão (6.13) pode ser utilizada para obtenção das matrizes coeficiente $\boldsymbol{\Theta_j}$ a partir dos coeficientes de correlação cruzada $\boldsymbol{\Gamma(l)}$'s de um processo MA(q).

Representação de um processo VARMA na forma VAR(1)

(Lütkepohl, 2007) propõe de modelo VARMA numa representação na forma VAR(1), que poderá ser bastante útil para predição ou, fundamentalmente, para representação em espaço de estados.

Suponha-se que Z_t tem uma representação VARMA(p, q) standard (6.7)

$$\boldsymbol{(Z_t - \mu) - \sum_{j=1}^{p} \Phi_j(Z_{t-j} - \mu) = \varepsilon_t - \sum_{j=1}^{q} \Theta_j\, \varepsilon_{t-j}}$$

Ou, na notação utilizada por (Lütkepohl, 2007) cujas equivalências são obvias

$$\boldsymbol{y_t = \nu + \sum_{j=1}^{p} A_j y_{t-j} + \sum_{j=1}^{q} M_j\, u_{t-j} + u_t}$$

Por simplicidade, assume-se média nula e consequentemente, $\boldsymbol{\nu = 0}$. Defina-se

$$\boldsymbol{Y_t :=} \underbrace{\begin{bmatrix} y_t \\ \vdots \\ y_{t-p+1} \\ u_t \\ \vdots \\ u_{t-q+1} \end{bmatrix}}_{(k(p+q)\times 1)} \qquad \boldsymbol{U_t :=} \begin{matrix} \left.\begin{bmatrix} u_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}\right\}(k_p \times 1) \\ \left.\begin{bmatrix} u_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}\right\}(k_q \times 1) \end{matrix}$$

e

$$\mathbf{A} \coloneqq \begin{bmatrix} \boldsymbol{A_{11}} & \boldsymbol{A_{12}} \\ \boldsymbol{A_{21}} & \boldsymbol{A_{22}} \end{bmatrix} \quad \left(\boldsymbol{k(p+q)} \times \boldsymbol{k(p+q)}\right)$$

onde

$$\boldsymbol{A_{11}} \coloneqq \underbrace{\begin{bmatrix} \boldsymbol{A_1} & \cdots & \boldsymbol{A_{p-1}} & \boldsymbol{A_p} \\ \boldsymbol{I_k} & & \boldsymbol{0} & \boldsymbol{0} \\ & \ddots & & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{I_k} & \boldsymbol{0} \end{bmatrix}}_{\boldsymbol{kp \times kp}}$$

$$\boldsymbol{A_{12}} \coloneqq \underbrace{\begin{bmatrix} \boldsymbol{M_1} & \cdots & \boldsymbol{M_{q-1}} & \boldsymbol{M_q} \\ \boldsymbol{0} & & \boldsymbol{0} & \boldsymbol{0} \\ & \ddots & & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}}_{\boldsymbol{kp \times kp}}$$

$$\boldsymbol{A_{21}} \coloneqq \underbrace{\boldsymbol{0}}_{\boldsymbol{kp \times kp}}$$

$$\boldsymbol{A_{11}} \coloneqq \underbrace{\begin{bmatrix} \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{I_k} & & \boldsymbol{0} & \boldsymbol{0} \\ & \ddots & & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{I_k} & \boldsymbol{0} \end{bmatrix}}_{\boldsymbol{kp \times kp}}$$

Com esta notação, obtém-se a representação VAR(1) de um processo VARMA(*p,q*)

$$\boldsymbol{Y_t = AY_{t-1} + U_t}$$

6.1.2.2 Aspectos da não unicidade para os modelos vectoriais ARMA

Para um modelo misto vector ARMA(*p,q*) dado por (6.7) com $p > 0$ e $q > 0$ especificamente, são necessárias certas condições nos operadores matriciais $\boldsymbol{\Phi(\mathcal{B})}$ e $\boldsymbol{\Theta(\mathcal{B})}$ para se garantir uma representação única do modelo ARMA e a identificabilidade de parâmetros. De notar que esta situação não se verifica nos modelos puros AR(*p*) e MA(*q*), mas ao evitar-se o uso de modelos mistos ARMA(*p,q*), pode por vezes ser-se conduzido à utilização de modelos puros não muito parcimoniosos no numero de parâmetros necessário à representação do processo. Especificamente na situação multivariada, é possível que para duas representações ARMA(*p, q*), $\boldsymbol{\Phi(\mathcal{B})Z_t = \Theta(\mathcal{B})\varepsilon_t}$ e $\boldsymbol{\Phi_*(\mathcal{B})Z_t = \Theta_*(\mathcal{B})\varepsilon_t}$ com diferentes coeficientes matriciais, dêem os mesmos coeficientes $\boldsymbol{\Psi_j}$ na representação MA infinita $\boldsymbol{Z_t = \Psi(\mathcal{B})\varepsilon_t = \sum_{j=0}^{\infty} \Psi_j\, \varepsilon_{t-j}}$ do processo, ou seja

$$\boldsymbol{\Psi(\mathcal{B}) = \Phi(\mathcal{B})^{-1}\Theta(\mathcal{B}) = \Phi_*(\mathcal{B})^{-1}\Theta_*(\mathcal{B})}$$

Consequentemente, representações diferentes dão origem à mesma estrutura matricial de covariância $\boldsymbol{\Gamma(l)}$ do processo. Duas representações de um modelo ARMA(*p, q*) com estas

propriedades dizem-se *observacionalmente* equivalentes. (Box, Jenkins, & Reinsel, 2008) dá o exemplo de um modelo ARMA(1,1) bivariado $(\boldsymbol{I}-\boldsymbol{\Phi}_*\boldsymbol{\mathcal{B}})\boldsymbol{Z_t}=(\boldsymbol{I}-\boldsymbol{\Theta}_*\boldsymbol{\mathcal{B}})\boldsymbol{\varepsilon_t}$ com a forma

$$\boldsymbol{\Phi}_* = \begin{bmatrix}\boldsymbol{0} & \boldsymbol{\alpha}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix} \qquad \boldsymbol{\Theta}_* = \begin{bmatrix}\boldsymbol{0} & \boldsymbol{\beta}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}$$

Que é observavelmente equivalente a um modelo AR(1) $(\boldsymbol{I}-\boldsymbol{\Phi\mathcal{B}})\boldsymbol{Z_t}=\boldsymbol{\varepsilon_t}$ e a um modelo MA(1) $\boldsymbol{Z_t}=(\boldsymbol{I}-\boldsymbol{\Theta\mathcal{B}})\boldsymbol{\varepsilon_t}$ com

$$\boldsymbol{\Phi} \equiv -\boldsymbol{\Theta} = \begin{bmatrix}\boldsymbol{0} & (\boldsymbol{\alpha}-\boldsymbol{\beta})\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}$$

Este problema é similar, mas mais complexo, que o problema da sobreparametrização e redundância de parâmetros que existe no caso ARMA univariado.

Os parâmetros de um modelo ARMA(p, q) são identificáveis se os parâmetros $\boldsymbol{\Phi_j}$ e $\boldsymbol{\Theta_j}$ são unicamente determinados pelas matrizes da resposta ao impulso $\boldsymbol{\Psi_j}$ do operador $\boldsymbol{\Psi}(\boldsymbol{\mathcal{B}})$ da representação MA infinita (única) do processo, ou seja, se não existe outro par de operadores ARMA(p,q) $\left(\boldsymbol{\Phi}_*(\boldsymbol{\mathcal{B}}),\boldsymbol{\Theta}_*(\boldsymbol{\mathcal{B}})\right)$ que seja *observacionalmente* equivalente a $\left(\boldsymbol{\Phi}(\boldsymbol{\mathcal{B}}),\boldsymbol{\Theta}(\boldsymbol{\mathcal{B}})\right)$ (Reinsel, 1997).

Para além das condições de invertibilidade e estacionaridade,

i. Todas a s raízes de $det\{\boldsymbol{\Phi}(\boldsymbol{\mathcal{B}})\}=0$ e todas as raízes de $det\{\boldsymbol{\Theta}(\boldsymbol{\mathcal{B}})\}=0$ têm de ser maiores que zero em valor absoluto

As duas condições seguintes são suficientes para a identificabilidade dos parâmetros num modelo ARMA(*p,q*) (Reinsel, 1997)

ii. As matrizes $\boldsymbol{\Phi}(\boldsymbol{\mathcal{B}})$ e $\boldsymbol{\Theta}(\boldsymbol{\mathcal{B}})$ não tem factores comuns, a não ser factores unimodulares, isto é, se $\boldsymbol{\Phi}(\boldsymbol{\mathcal{B}})=\boldsymbol{U}(\boldsymbol{\mathcal{B}})\boldsymbol{\Phi_1}(\boldsymbol{\mathcal{B}})$ e $\boldsymbol{\Theta}(\boldsymbol{\mathcal{B}})=\boldsymbol{U}(\boldsymbol{\mathcal{B}})\boldsymbol{\Theta_1}(\boldsymbol{\mathcal{B}})$, então o factor comum $\boldsymbol{U}(\boldsymbol{\mathcal{B}})$ tem de ser unimodular, isto é $det\{\boldsymbol{U}(\boldsymbol{\mathcal{B}})\}$ é uma constante (diferente de zero), e

iii. Com q menor possível p menor possível para este q, a matriz $\left[\boldsymbol{\Phi_p},\boldsymbol{\Theta_q}\right]$ tem que ter rank completo

Quando se verifica a propriedade (ii), os operadores $\boldsymbol{\Phi}(\boldsymbol{\mathcal{B}})$ e $\boldsymbol{\Theta}(\boldsymbol{\mathcal{B}})$ dizem-se *left-coprime*, e a representação $\boldsymbol{\Psi}(\boldsymbol{\mathcal{B}})=\boldsymbol{\Phi}(\boldsymbol{\mathcal{B}})^{-1}\boldsymbol{\Theta}(\boldsymbol{\mathcal{B}})$ diz-se ser irredutível.

Uma abordagem particular para identificação, é representar o modelo ARMA numa certa forma "canónica" a partir da qual existe um único modelo representativo dessa forma para cada classe de modelos ARMA observacionalmente equivalentes. Essa abordagem será discutida na secção 6.1.2.3 em relação à forma canónica escalão (*echelon*).

Modelos vectoriais ARMA não estacionários

Na prática, muitos processos têm comportamento não estacionário, frequentemente de natureza homogénea, como deriva ou tendência a nível local, mas que no global apresentam um padrão homogéneo. Frequentemente, por exemplo num modelo ARIMA univariado, a não estacionaridade pode ser retirada diferenciando as séries.

Para os modelos vectoriais ARMA, as condições para a estacionaridade exigem que todas as raízes de $det\{\Phi(\mathcal{B})\} = 0$ sejam maiores que um em valor absoluto. Generalizando para processos não estacionários, mas não explosivos, pode-se considerar uma forma geral de modelos vectoriais ARMA, $\boldsymbol{\Phi(\mathcal{B})Z_t} =, \boldsymbol{\Theta(\mathcal{B})\varepsilon_t}$, onde alguns valores de $det\{\Phi(\mathcal{B})\} = 0$ podem ter o valor absoluto igual a um. Mais especificamente, devido ao papel proeminente do operador $(1 - \mathcal{B})$ nos modelos univariados, pode-se permitir apenas algumas raízes unitárias, sendo as restantes superiores a um em valor absoluto.

Considere-se primeiramente modelos da forma

$$\boldsymbol{\Phi_1(\mathcal{B})}\mathbf{D}(\mathcal{B})\boldsymbol{Z_t} = \boldsymbol{\Theta(\mathcal{B})\varepsilon_t} \tag{6.14}$$

Onde $\mathbf{D}(\mathcal{B}) = \mathbf{diag}\left[(\mathbf{1} - \mathcal{B})^{d_1}, (\mathbf{1} - \mathcal{B})^{d_2}, \ldots\ , (\mathbf{1} - \mathcal{B})^{d_k}\right]$ é uma matriz diagonal, $d_1, d_2, \ \ldots\ , d_k$ são inteiros positivos, e $\boldsymbol{det\{\Phi_1(\mathcal{B})\} = 0}$ tem todas a raízes maiores que um em valor absoluto. Este modelo que é referido como modelo vectorial ARIMA, apenas estabelece que após a diferenciação individualmente de cada série Z_{it} de um apropriado numero d_i de vezes para a reduzir a uma série estacionária, a série vectorial resultante $\boldsymbol{W_t} = \mathbf{D}(\mathcal{B})\boldsymbol{Z_t}$ será um processo vectorial ARMA(*p,q*) estacionário. Como se verá seguidamente, este modelo quase não é tão geral como o modelo $\boldsymbol{\Phi_1(\mathcal{B})Z_t} = \boldsymbol{\Theta(\mathcal{B})\varepsilon_t}$, onde são permitidas algumas raízes de $\boldsymbol{det\{\Phi_1(\mathcal{B})\} = 0}$ com valor absoluto igual a um. A utilização destes modelos em situações não apropriadas poderá conduzir a sobre diferenciação e consequentemente à não invertibilidade do operador MA.

O aspecto da não estacionaridade (raízes unitárias) de um processo vectorial $\boldsymbol{Z_t}$ torna-se assim mais complicado comparativamente ao caso univariado, devido em parte à possibilidade de cointegração entre as séries componentes Z_{it} de um processo vectorial não estacionário. Por exemplo, existe a possibilidade de uma série componente Z_{it} ser não estacionária, com a série das suas primeiras diferenças $(1 - \mathcal{B})Z_{it}$ estacionária (neste caso Z_{it} dir-se-á cointegrada de ordem um), mas tal que ceras combinações lineares $Z_{it} = \boldsymbol{b_i'Z_t}$ serão estacionárias. Nestes casos o processo $\boldsymbol{Z_t}$ diz-se cointegrado com vector de cointegração $\boldsymbol{b_i}$. Uma interpretação de cointegração, particularmente relacionada com modelos económicos, é que as séries componentes individuais partilham algumas componentes não estacionaria comuns ou "tendências comuns", e consequentemente, tem tendência a ter alguns movimentos similares no seu comportamento a longo prazo. Uma estrutura especifica de modelo ARMA não estacionário em que ocorre cointegração é o modelo $\boldsymbol{\Phi(\mathcal{B})Z_t} = \boldsymbol{\Theta(\mathcal{B})\varepsilon_t}$, onde $\boldsymbol{det\{\Phi_1(\mathcal{B})\} = 0}$ tem $d < k$ raízes unitárias e todas as restantes maiores que um em valor absoluto, e em que a matriz $\boldsymbol{\Phi(1)}$ tem rank igual a $r = k - d$. Deste modo, pode estabelecer-se que existem r vectores linearmente independentes $\boldsymbol{b_i}$ tal que $\boldsymbol{b_i'Y_t}$ tem comportamento estacionário. Diz-se então que $\boldsymbol{Y_t}$ tem cointegração rank r.

6.1.2.3 Forma canónica escalão do modelo VARMA

No campo da modelação, o objectivo principal deste trabalho é determinar uma função de transferência para um processo na forma VARMA estandardizada (6.7) para ser utilizada no algoritmo de predição e/ou controlo. Contudo, o esforço de modelação, no sentido de se obter o modelo mais próximo possível do comportamento do processo, poderá ter objectivos diferentes como a filtragem do processo com o objectivo de se aplicar controlo estatístico

multivariado, ou apenas para previsão do comportamento futuro do processo. Neste sentido, e também porque o objectivo do trabalho passa por obter um modelo do processo que apresente o melhor desempenho possível, apresenta-se uma abordagem muito ligeira a outras formas representativas que poderão revelar-se mais úteis para especificação dos modelos dos processos que a forma estandardizada.

Um vector ARMA(*p,q*) pode ser escrito na seguinte forma equivalente

$$\boldsymbol{\Phi}_0^{\#}(\boldsymbol{Z}_t - \boldsymbol{\mu}) - \sum_{j=1}^{p} \boldsymbol{\Phi}_j^{\#}\left(\boldsymbol{Z}_{t-j} - \boldsymbol{\mu}\right) = \boldsymbol{\Theta}_0^{\#}\boldsymbol{\varepsilon}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j^{\#}\,\boldsymbol{\varepsilon}_{t-j} \qquad \textbf{(6.15)}$$

onde $\boldsymbol{\Phi}_0^{\#}$ é uma matriz não singular arbitrária, com $\boldsymbol{\Theta}_0^{\#} = \boldsymbol{\Phi}_0^{\#}$, $\boldsymbol{\Phi}_j^{\#} = \boldsymbol{\Phi}_0^{\#}\boldsymbol{\Phi}_j$, e $\boldsymbol{\Theta}_j^{\#} = \boldsymbol{\Phi}_0^{\#}\boldsymbol{\Theta}_j$. Para se obter uma identificabilidade única dos parâmetros do modelo na forma (6.15), torna-se necessário restringir (normalizar) a forma da matriz $\boldsymbol{\Phi}_0^{\#}$, que deverá ser, no mínimo, triangular inferior com uns (1's) na diagonal.

Para especificar a forma canónica, para além dos habituais indicies *p* e *q*, é necessário determinar outro conjunto de índices, chamados índices de *Kronecker* $K_1, K_2, \ldots, K_k$, que irão corresponder à ordem dos polinómios em $\boldsymbol{\mathcal{B}}$ de cada uma das k variáveis que compõem o vector Z. A forma canónica de um modelo ARMA é determinada como sendo a representação (6.15) tal que $[\Phi^{\#}(\boldsymbol{\mathcal{B}}),\ \Theta^{\#}(\boldsymbol{\mathcal{B}})]$ tem o menor grau de linha possível, em que K_i é o grau da linha i de $[\Phi^{\#}(\boldsymbol{\mathcal{B}}),\ \Theta^{\#}(\boldsymbol{\mathcal{B}})]$, para i= $1, \ldots, k$, em que o grau do modelo (*p, q*) vai ser determinado por $p = q = max\{K_1, K_2, \ldots, K_k\}$. A especificação destes índices, conhecidos como índices de *Kronecker*, que são únicos para qualquer classe equivalente de modelos ARMA, e determinam o modelo (6.15) de forma única, no qual os parâmetros desconhecidos são identificados de forma única.

Grau McMillan de um processo vectorial ARMA

Para qualquer processo vectorial $\{\boldsymbol{Z}_t\}$ com matrizes de covariância $\Gamma(l)$, define-se a matriz bloco de dimensão infinita, denominada como Matriz de Hankel de covariância por

$$\mathbf{H} = \begin{bmatrix} \boldsymbol{\Gamma}(1)' & \boldsymbol{\Gamma}(2)' & \boldsymbol{\Gamma}(3)' & \cdots \\ \boldsymbol{\Gamma}(2)' & \boldsymbol{\Gamma}(3)' & \boldsymbol{\Gamma}(4)' & \cdots \\ \boldsymbol{\Gamma}(3)' & \boldsymbol{\Gamma}(4)' & \boldsymbol{\Gamma}(5)' & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

O *grau de McMillam* ***M*** do processo $\{\boldsymbol{Z}_t\}$ é definido pelo rank da matriz **H**. O processo $\{\boldsymbol{Z}_t\}$ segue um modelo ARMA finito se e só se o rank da matriz **H** é finito. Para um processo ARMA(p,q), a relação de momentos (6.7) resulta em

$$\boldsymbol{\Gamma}(\boldsymbol{l})' - \sum_{j=1}^{p} \boldsymbol{\Phi}_j \boldsymbol{\Gamma}(\boldsymbol{l}-\boldsymbol{j})' = \mathbf{0}$$

para $l > q$. A partir daqui pode-se verificar que o rank de **H** (o grau de McMillan) satisfará a relação $M \leq ks$, onde $s = Max\{p, q\}$, consequentemente verificar-se-á por exemplo a seguinte relação

$$[-\boldsymbol{\Phi_s},\ -\boldsymbol{\Phi_{s-1}},\ \ldots,\ -\boldsymbol{\Phi_1},\ \mathbf{I},\ \mathbf{0},\ \mathbf{0},\ \cdots]\mathbf{H} = \mathbf{0}$$

com $\boldsymbol{\Phi}_\mathbf{j} = \mathbf{0}$ para $\boldsymbol{j} > \boldsymbol{p}$. Consequentemente, todas a linhas constituídas por blocos $k \times k$ de **H** para além do *s*-ésimo bloco linha serão linearmente dependentes das linhas bloco precedentes. Contudo, o grau de *McMillan* de um vector ARMA(*p,q*), $\boldsymbol{M} = \sum_{i=1}^{k} \boldsymbol{K_i} = \boldsymbol{K_1}, \boldsymbol{K_2}, \ldots, \boldsymbol{K_k}$, poderá ser consideravelmente inferior a ks devido às deficiências de rank nos coeficientes matriciais MA e AR.

O grau de McMillan ***M*** é interpretado como o número de combinações lineares linearmente independentes dos vectores passados e presente $\boldsymbol{Z_t}, \boldsymbol{Z_t}, \ldots$ que são para uma predição óptima para todos os vectores futuros dentro da estrutura ARMA (Reinsel, 1997).

Forma canónica implícita pelos índices de *Kronecker*

A exemplo do que foi feito para a expressão (6.6), as matrizes de covariância cruzada $\Gamma(l)$ do processo definido por (6.15), serão dadas por

$$\boldsymbol{\Phi}_\mathbf{0}^{\#}\boldsymbol{\Gamma}(\boldsymbol{l})' - \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\Phi}_\mathbf{j}^{\#}\boldsymbol{\Gamma}(\boldsymbol{l}-\boldsymbol{j})' = -\sum_{\boldsymbol{j=l}}^{\mathbf{q}} \boldsymbol{\Theta}_\mathbf{j}^{\#}\,\boldsymbol{\Sigma\Psi'}_{\boldsymbol{j-l}}$$

Se $\phi_j(i)'$ for a *i*-ésima linha de $\boldsymbol{\Phi}_\mathbf{j}^{\#}$, então o *i*-ésimo índice de *Kronecker* que será k_i implica a dependência linear nas linhas da matriz de *Hankel* da forma

$$\phi_0(i)'\,\boldsymbol{\Gamma}(\boldsymbol{l})' - \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \phi_j(i)'\,\boldsymbol{\Gamma}(\boldsymbol{l}-\boldsymbol{j})' = \mathbf{0} \qquad para\ \ l \geq K_i + 1 \tag{6.16}$$

Que pode ser na forma $b'_i H = 0'$ com $b'_i = \left(-\phi_{K_i}(i)', \ldots, -\phi_1(i)',\ \phi_0(i)',\ 0', \ldots\right)$. De notar que, por definição do *i*-ésimo índice de *Kronecker* K_i, o vector $\phi_0(i)'$ na expressão anterior deve ser construído de modo que a componente índice *i* seja 1 e que as componentes com índice superior a *i* sejam nulas. Consequentemente, o índice de *Kronecker* igual a K_i, implica, em particular, que a representação do modelo ARMA da forma (6.15) deve ser construída de forma a que *i*-ésima linha das matrizes $\boldsymbol{\Phi}_\mathbf{j}^{\#}$ e $\boldsymbol{\Theta}_\mathbf{j}^{\#}$ sejam zero para $j > K_i$.

Adicionalmente, a expressão (6.16) implica que certos elementos da linha *i* das matrizes $\boldsymbol{\Phi}_\mathbf{j}^{\#}$, para $j \leq K_i$ sejam nulos. Especificamente, o *l*-ésimo elemento da *i*-ésima linha de $\phi_0(i)'$ deve ser zero sempre que $j + K_l \leq K_i$, porque para $K_l \leq K_i$ as linhas $k(K_l + j) + l,\ j = 0, \ldots, (K_i - K_l)$, da matriz de Hankel são todas linearmente independentes das linhas anteriores. Consequentemente, o elemento índice (i, l) do operador AR

$$\boldsymbol{\Phi}^{\#}(\boldsymbol{\mathcal{B}}) = \boldsymbol{\Phi}_\mathbf{0}^{\#} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\Phi}_\mathbf{j}^{\#}\boldsymbol{\mathcal{B}}^{\boldsymbol{j}}$$

em (6.15) pode ser especificado para ter coeficientes diferentes de zero apenas para as lags $j = K_i - K_{il} + 1, \ldots, K_i$, com coeficientes iguais a zero para qualquer *lag* abaixo de *j* (quando $i \neq j$), onde se define

$$K_{il} = \begin{cases} min(K_i + 1, K_l) & para\ \ i > l \\ min(K_i, K_l) & para\ i \geq l \end{cases}$$

(tal que sempre que $K_l \leq K_i$ se tenha $K_{il} = K_l$). Deste modo, o correspondente numero de parâmetros de AR desconhecidos no elemento (i, l) de $\boldsymbol{\Phi}^{\#}(\boldsymbol{\mathcal{B}})$ é igual a K_{il}. Assim, o operador AR $\boldsymbol{\Phi}^{\#}(\boldsymbol{\mathcal{B}})$ no modelo (6.15) pode ser especificado de forma a que o numero total de parâmetros desconhecidos de $\boldsymbol{\Phi}^{\#}(\boldsymbol{\mathcal{B}})$ é igual a $\sum_{i=1}^{k}\sum_{l=1}^{k} K_{il} = M + \sum\sum_{i\neq l}^{k} K_{il}$, enquanto que o numero de parâmetros desconhecidos no operador MA $\boldsymbol{\Theta}^{\#}(\boldsymbol{\mathcal{B}})$ em (6.15), excluindo os parâmetros referentes a $\boldsymbol{\Theta}_{\mathbf{0}}^{\#} = \boldsymbol{\Phi}_{\mathbf{0}}^{\#}$ é igual a $\sum_{i=1}^{k} k\, K_i = kM$ (Reinsel, 1997).

Forma de rank reduzido do modelo VARMA implícita pelos índices de Kronecker

Multiplicando o modelo ARMA na forma canónica (6.15) por $\boldsymbol{\Phi}_{\mathbf{0}}^{\#\,-\mathbf{1}}$, obtém-se o modelo vectorial ARMA na forma standard

$$\begin{aligned} \boldsymbol{Z_t} &= \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\Phi}_{\mathbf{0}}^{\#\,-\mathbf{1}} \boldsymbol{\Phi}_{\mathbf{j}}^{\#} \boldsymbol{Z_{t-j}} + \boldsymbol{\varepsilon_t} - \sum_{\mathbf{j=1}}^{\mathbf{q}} \boldsymbol{\Phi}_{\mathbf{0}}^{\#\,-\mathbf{1}} \boldsymbol{\Theta}_{\mathbf{j}}^{\#}\, \boldsymbol{\varepsilon_{t-j}} \\ &\equiv \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\Phi_j} \boldsymbol{Z_{t-j}} + \boldsymbol{\varepsilon_t} - \sum_{\mathbf{j=1}}^{\mathbf{q}} \boldsymbol{\Theta_j}\, \boldsymbol{\varepsilon_{t-j}} \end{aligned} \qquad \textbf{(6.17)}$$

Nesta forma standard do modelo VARMA, as matrizes coeficiente AR e MA, respectivamente $\boldsymbol{\Phi_j} = \boldsymbol{\Phi}_{\mathbf{0}}^{\#\,-\mathbf{1}} \boldsymbol{\Phi}_{\mathbf{j}}^{\#}$ e $\boldsymbol{\Theta_j} = \boldsymbol{\Phi}_{\mathbf{0}}^{\#\,-\mathbf{1}} \boldsymbol{\Theta}_{\mathbf{j}}^{\#}$, podem adquirir uma forma especial de rank reduzido. Seja r_j, com $j = 1, \ldots, p$, o rank das matrizes $(\boldsymbol{\Phi_j},\ \boldsymbol{\Theta_j}) = \left(\boldsymbol{\Phi}_{\mathbf{0}}^{\#\,-\mathbf{1}} \boldsymbol{\Phi}_{\mathbf{j}}^{\#},\ \boldsymbol{\Phi}_{\mathbf{0}}^{\#\,-\mathbf{1}} \boldsymbol{\Theta}_{\mathbf{j}}^{\#}\right)$. Tem-se então que o rank r_j é igual ao numero de índices de *Kronecker* que são iguais ou maiores que a lag j, ou seja, se no conjunto dos k índices de *Kronecker* existirem *n* índices com valor igual ou superior a *j*, então o valor de r_j coincidirá com o valor *n*. Consequentemente, as matrizes $(\boldsymbol{\Phi}_{\mathbf{j}}^{\#},\ \boldsymbol{\Theta}_{\mathbf{j}}^{\#})$ terão uma linha só com zeros para cada índice de *Kronecker* em que $j > K_i$. Deste modo, o rank r_j vai decrescendo à medida que j aumenta. Denotando $D_j^{'}$ como sendo a submatriz $r_j \times k$ da matriz identidade $k \times k$ que selecciona as linhas não nulas de $(\boldsymbol{\Phi}_{\mathbf{j}}^{\#},\ \boldsymbol{\Theta}_{\mathbf{j}}^{\#})$, tal que $(\mathbf{B_j},\ \mathbf{C_j}) \equiv \boldsymbol{D_j^{'}}(\boldsymbol{\Phi}_{\mathbf{j}}^{\#},\ \boldsymbol{\Theta}_{\mathbf{j}}^{\#})$ são composta apenas pelas r_j linhas não nulas de $(\boldsymbol{\Phi}_{\mathbf{j}}^{\#},\ \boldsymbol{\Theta}_{\mathbf{j}}^{\#})$. Então tem-se que

$$(\boldsymbol{\Phi_j},\ \boldsymbol{\Theta_j}) = \boldsymbol{\Phi}_{\mathbf{0}}^{\#\,-\mathbf{1}} \boldsymbol{D_j}(\boldsymbol{B_j},\ \boldsymbol{C_j}) \equiv \boldsymbol{A_j}(\boldsymbol{B_j},\ \boldsymbol{C_j})$$

onde $\boldsymbol{A_j} = \boldsymbol{\Phi}_{\mathbf{0}}^{\#\,-\mathbf{1}} \boldsymbol{D_j}$ é uma matriz $k \times r_j$. Com esta expressão obtém-se uma factorização de rank reduzido das matrizes coeficiente $(\boldsymbol{\Phi_j},\ \boldsymbol{\Theta_j})$ na forma standard (6.17). Assim, o modelo vectorial ARMA na forma standard pode ser escrito como

$$\boldsymbol{Z_t} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{A_j B_j Z_{t-j}} + \boldsymbol{\varepsilon_t} - \sum_{\mathbf{j=1}}^{\mathbf{q}} \boldsymbol{A_j C_j}\, \boldsymbol{\varepsilon_{t-j}} \qquad \textbf{(6.18)}$$

O modelo vectorial ARMA desta forma pode ser referido como representação de rank reduzido encadeado do modelo ARMA (*nested reduzed-rank ARMA model representation*) (Reinsel, 1997).

6.1.2.4 Estrutura de Correlação Canónica para séries temporais ARMA

Como descrito na secção 4.2.6.2 - Método das correlações canónicas, a análise de correlação canónica consiste em determinar, para dois conjuntos de variáveis, $X_1 = (x_{11},\ x_{12},\ \dots\ , x_{1k_1})'$ e $Y_2 = (x_{21},\ x_{22},\ \dots\ , x_{2k_2})'$, de dimensões k_1 e k_2 respectivamente (assume-se que $k_1 \le k_2$), combinações lineares $U_i = a_i' X_1$ e $V_i = b_i' X_2,\ \ i = 1,\ \dots\ , k$ e as correspondentes correlações $\rho_i = corr[U_i,\ V_i]$ com $\rho_1 \ge \rho_2 \ge \cdots \ge \rho_{k_1} \ge 0$, tal que tal que os U_i e os V_j estão mutuamente não correlacionados para $i \ne j$. U_1 e V_1 tem a máxima correlação possível ρ_1 entre todas as combinações lineares de X_1 e X_2, U_2 e V_2 tem a máxima correlação possível ρ_2 entre todas as combinações lineares de X_1 e X_2 que não estão correlacionadas com U_1 e V_1, e assim sucessivamente.

As estimativas amostrais de correlações canónicas de populações são construídas de forma óbvia. Suponha-se que $X_1 = (X_{1t}',\ X_{2t}')',\ t = 1, \dots\ , T$, é uma amostra aleatória de T observações vectoriais da distribuição $X = (X_1',\ X_2')'$. Então, a matriz de covariância amostral é dada por $\widehat{\Omega} = T^{-1}\sum_{i=1}^{t}(X_t - \bar{X})(X_t - \bar{X})'$, onde $\bar{X} = T^{-1}\sum_{i=1}^{t} X_t$ é obviamente o vector da média amostral, e seja $\widehat{\Omega}$ particionada de obvia com $\widehat{\Omega}_{\text{ij}} = T^{-1}\sum_{i=1}^{t}(X_{it} - \bar{X}_i)\left(X_{jt} - \bar{X}_j\right)'$ para $i = 1, 2$. Então as *correlações canónicas amostrais* são os valores $\hat{\rho}_1 > \hat{\rho}_2 > \cdots > \hat{\rho}_{k_1} > 0$ tal que $\hat{\rho}_i^2$ são os valores próprios ordenados da matriz $\widehat{\Omega}_{11}^{-1}\widehat{\Omega}_{12}\widehat{\Omega}_{22}^{-1}\widehat{\Omega}_{21}$, tal que os $\hat{\rho}_i^2$ satisfazem a relação

$$\left(\widehat{\boldsymbol{\rho}}_{\boldsymbol{i}}^{\boldsymbol{2}}\boldsymbol{I} - \widehat{\boldsymbol{\Omega}}_{\boldsymbol{11}}^{\boldsymbol{-1}}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{12}}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{22}}^{\boldsymbol{-1}}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{21}}\right)\widehat{\boldsymbol{a}}_{\mathbf{i}} = \mathbf{0}, \qquad \mathbf{i} = \mathbf{1}, \dots\ , \boldsymbol{k_1} \tag{6.19}$$

onde os $\widehat{\boldsymbol{a}}_{\mathbf{i}}$ são os respectivos vectores próprios. As correlações canónicas parciais amostrais entre $\{X_{1t}\}$ e $\{X_{2t}\}$, dado $\{X_{3t}\}$ são definidas de forma similar baseado na amostra $X_t = (X_{1t}',\ X_{2t}',\ X_{3t}')',\ t = 1, \dots\ , T$, seguindo o procedimento descrito na secção 4.2.6.2 - Método das correlações canónicas.

Na análise de correlação canónica, está-se frequentemente interessado nas combinações lineares de X_1 e X_2 que possuam uma forte correlação e, talvez ainda mais importante, quando é que algumas das correlações canónicas entre X_1 e X_2 são (essencialmente) nulas, uma vez que esse facto pode conduzir a uma redução substancial da dimensão do estudo das relações entre os dois conjuntos de variáveis. Uma propriedade importante é que se existirem (no mínimo) s $(s \le k_1)$ combinações lineares linearmente independentes de X_1 que são completamente não correlacionadas com X_2, seja por exemplo $U = A'X_1$ tal que $Cov(X_1,\ U) = \Omega_{21}\mathrm{A} = 0$, então existem no mínimo s correlações canónicas entre X_1 e X_2 que são nulas. O facto de se ter $\Omega_{21}\mathrm{A} = 0$ implica as s colunas linearmente independentes de A satisfazem a relação (4.105), $[\rho^2 I - \Omega_{11}^{-1}\Omega_{12}\Omega_{22}^{-1}\Omega_{21}]A = 0$, para $\rho = 0$, e consequentemente existem (no mínimo) s valores próprios em (6.19). Com efeito, o numero s de correlações canónicas nulas é igual a $k_1 - r\ \ (s = k_1 - r)$, onde $r = rank(\Omega_{21})$. Baseado na amostra de $\boldsymbol{T}$ observações vectoriais, o teste de hipótese de que as ultimas $k_1 - r$ correlações canónicas populacionais entre X_1 e X_2 são nulas é dado por

$$-\mathbf{2log}(\boldsymbol{\Lambda}) = \boldsymbol{T}\sum_{\boldsymbol{i=r+1}}^{\boldsymbol{k_1}}\mathbf{log}\left(\mathbf{1} - \widehat{\boldsymbol{\rho}}_{\boldsymbol{i}}^{\mathbf{2}}\right) \tag{6.20}$$

Este teste segue uma distribuição $\boldsymbol{\chi^2}$ com $(k_1 - r\ )(k_2 - r\ )$ graus de liberdade (Reinsel, 1997).

6.1.2.5 Modelo vectorial autoregressivo e matrizes de autoregressão parcial

Para um processo vectorial estacionário $\{\boldsymbol{Z_t}\}$ com matriz de covariância $\boldsymbol{\Gamma(l)}$, poderá ser por vezes conveniente, especialmente na fase de determinação da ordem do processo, ou para propósitos de predição, aproximar o modelo do processo a um modelo vectorial autoregressivo, sendo o processo um AR puro ou não. Assumindo-se um processo com media nula, para uma determinada ordem m, pode-se determinar os coeficientes matriciais $\Phi_{1\mathrm{m}}, \Phi_{2\mathrm{m}}, \ldots, \Phi_{\mathrm{mm}}$ numa aproximação a um modelo autoregressivo que minimize a quantidade

$$\boldsymbol{tr}\left\{\boldsymbol{E}\left[\left(\boldsymbol{Z_t}-\sum_{\boldsymbol{j=1}}^{\boldsymbol{m}}\boldsymbol{\Phi_{jm}Z_{t-j}}\right)\left(\boldsymbol{Z_t}-\sum_{\boldsymbol{j=1}}^{\boldsymbol{m}}\boldsymbol{\Phi_{jm}Z_{t-j}}\right)'\right]\right\}$$

O valor esperado envolvido na expressão acima pode ser expresso como (Reinsel, 1997)

$$\boldsymbol{E}\left[\left(\boldsymbol{Z_t}-\boldsymbol{\Phi}'_{(m)}\boldsymbol{Z_{m,t-1}}\right)\left(\boldsymbol{Z_t}-\boldsymbol{\Phi}'_{(m)}\boldsymbol{Z_{m,t-1}}\right)'\right] = \boldsymbol{\Gamma(0)}-\boldsymbol{\Phi}'_{(m)}\boldsymbol{\Gamma}_{(m)}-\boldsymbol{\Gamma}'_{(m)}\boldsymbol{\Phi}_{(m)}+\boldsymbol{\Phi}'_{(m)}\boldsymbol{\Gamma_m}\boldsymbol{\Phi}_{(m)}$$

Onde

$$\boldsymbol{Z_{m,t-1}} = (\mathbf{Z}'_{\mathbf{t-1}}, \ldots, \mathbf{Z}'_{\mathbf{t-m}})$$
$$\boldsymbol{\Gamma_m} = \boldsymbol{E}\left(\boldsymbol{Z_{m,t-1}}\ \boldsymbol{Z}'_{\boldsymbol{m,t-1}}\right)$$
$$\boldsymbol{\Gamma}_{(m)} = \boldsymbol{E}\left(\boldsymbol{Z_{m,t-1}}\ \boldsymbol{Z}'_{\boldsymbol{t}}\right) = (\boldsymbol{\Gamma(0)}', \boldsymbol{\Gamma(1)}', \ldots, \boldsymbol{\Gamma(m)}')'$$
$$\boldsymbol{\Phi}_{(m)} = (\boldsymbol{\Phi_{1m}}, \boldsymbol{\Phi_{2m}}, \ldots, \boldsymbol{\Phi_{mm}},)$$

A minimização deste critério é um problema de regressão linear dos mínimos quadrados, multivariada normal. A matriz dos coeficientes $\boldsymbol{\Phi_{jm}}$ que minimizam o critério serão a solução da equação vectorial de *Yule-Walker* de ordem *m*,

$$\boldsymbol{\Phi}_{(m)} = \left\{\boldsymbol{E}\left(\boldsymbol{Z_{m,t-1}}\ \boldsymbol{Z}'_{\boldsymbol{m,t-1}}\right)\right\}^{-1}\boldsymbol{E}\left(\boldsymbol{Z_{m,t-1}}\ \boldsymbol{Z}'_{\boldsymbol{t}}\right) = \boldsymbol{\Gamma}_{\boldsymbol{m}}^{-1}\boldsymbol{\Gamma}_{(m)} \qquad \textbf{(6.21)}$$

Tal como no caso univariado, para $m = 1, 2, \ldots$, pode-se definir uma sequência de matrizes $\boldsymbol{\Phi_{mm}}$, a que se deu o nome de *matriz de autoregressão parcial* (*partial autoregression matrix*) de ordem (ou *lag*) *m*, como a solução para $\boldsymbol{\Phi_{mm}}$ das equações de Yule-Walker de ordem *m*, que resultam do ajustamento de um modelo AR de ordem *m* a uma série $\boldsymbol{Z_t}$ (Reinsel, 1997).

De forma similar ao caso univariado, segue-se que a sequência de matrizes autoregressivas parciais $\boldsymbol{\Phi_{mm}}$, de ordem $m = 1, 2, \ldots$, tem a importante qualidade característica de que, se o processo é de ordem p, então $\boldsymbol{\Phi_{pp}} = \boldsymbol{\Phi_p}$ e $\boldsymbol{\Phi_{mm}} = 0$ para $m > p$. Consequentemente, as matrizes $\boldsymbol{\Phi_{mm}}$ tem a propriedade de corte (*cutoff*) para um modelo AR(p), pelo esta propriedade pode verificar-se bastante útil na identificação e especificação de uma estrutura AR pura para um processo multivariado. Contudo, ao contrário do caso, os elementos das matrizes $\boldsymbol{\Phi_{mm}}$ não são correlações parciais como a função de autocorrelação parcial (PACF) $\boldsymbol{\phi_{mm}}$ no caso univariado.

6.1.3 Construção do modelo inicial e estimativa dos mínimos quadrados para modelos vectoriais ARMA

Nesta secção discute-se algumas técnicas para a especificação preliminar do modelo baseadas em estatísticas amostrais. São abordadas técnicas como estimativa dos mínimos quadrados e respectivos testes de hipótese associados. São ainda explorados alguns métodos adicionais para a especificação inicial e selecção de uma estrutura ARMA apropriada, tais como a utilização do método das correlações canónicas e critérios de selecção de modelos tais como AIC e BIC.

6.1.3.1 Estimativa do vector média e das matrizes de covariância e correlação

Dada uma amostra de uma serie temporal vectorial $\mathbf{Z_1}, \mathbf{Z_2}, \ldots, \mathbf{Z_T}$ de dimensão T, possivelmente, originada de um processo estacionário multivariado, com vector média $\boldsymbol{\mu} = \mathbf{E}(\mathbf{Z_t})$ e matrizes de autocovariância $\boldsymbol{\Gamma}(\boldsymbol{l}) = \boldsymbol{E}[(\mathbf{Z_t} - \boldsymbol{\mu})(\mathbf{Z_t} - \boldsymbol{\mu})']$ e matrizes de autocorrelação $\boldsymbol{\rho}(\boldsymbol{l})$, uma das ferramentas fundamentais para a determinação preliminar do modelo do processo é obter estimativas amostrais do vector $\boldsymbol{\mu}$ e especialmente das matrizes $\boldsymbol{\Gamma}(\boldsymbol{l})$ e $\boldsymbol{\rho}(\boldsymbol{l})$.

A estimativa preliminar de $\boldsymbol{\mu}$ é dada pelo vector média amostral

$$\hat{\boldsymbol{\mu}} = \overline{\boldsymbol{Z}} = \frac{\mathbf{1}}{\boldsymbol{T}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{T}} \mathbf{Z_t}$$

A estimativa amostral das matriz de covariância cruzada da *lag* l, $\boldsymbol{\Gamma}(\boldsymbol{l})$, é dada pela matriz de covariância amostral da *lag* l, definida por

$$\hat{\boldsymbol{\Gamma}}(\boldsymbol{l}) = \boldsymbol{C}(\boldsymbol{l}) = \frac{\mathbf{1}}{\boldsymbol{T}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{T-l}} (\boldsymbol{Z_t} - \overline{\boldsymbol{Z}})(\boldsymbol{Z_{t-l}} - \overline{\boldsymbol{Z}})', \qquad \boldsymbol{l} = \mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots \tag{6.22}$$

com $\hat{\boldsymbol{\Gamma}}(\boldsymbol{l}) = \hat{\boldsymbol{\Gamma}}(-\boldsymbol{l})'$ (ver secção 6.1.1.1). No caso particular da *lag* zero, tem-se que a matriz de covariância cruzada $\hat{\boldsymbol{\Gamma}}(\mathbf{0}) = \boldsymbol{C}(\mathbf{0}) = \boldsymbol{T}^{-1} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{Tl}} (\boldsymbol{Z_t} - \overline{\boldsymbol{Z}})(\boldsymbol{Z_t} - \overline{\boldsymbol{Z}})'$ corresponde à matriz de covariância amostral de $\boldsymbol{Z_t}$. Assim, o elemento (i, j) da matriz $\hat{\boldsymbol{\Gamma}}(\boldsymbol{l})$ é dado por

$$\hat{\boldsymbol{\gamma}}_{\boldsymbol{ij}}(\boldsymbol{l}) = \boldsymbol{c_{ij}}(\boldsymbol{l}) = \boldsymbol{T}^{-1} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{T-l}} (\boldsymbol{Z_{it}} - \overline{\boldsymbol{Z}}_{\boldsymbol{i}})\left(\boldsymbol{Z_{j,t-l}} - \overline{\boldsymbol{Z}}_{\boldsymbol{j}}\right)$$

As correlações cruzadas amostrais são definidas como

$$\hat{\boldsymbol{\rho}}_{\boldsymbol{ij}}(\boldsymbol{l}) = \boldsymbol{r_{ij}}(\boldsymbol{l}) = \frac{\boldsymbol{c_{ij}}(\boldsymbol{l})}{\left(\boldsymbol{c_{ii}}(\mathbf{0})\boldsymbol{c_{jj}}(\mathbf{0})\right)^{\mathbf{1/2}}}, \qquad \boldsymbol{i, j} = \mathbf{1}, \ldots, \boldsymbol{k} \tag{6.23}$$

Para uma serie estacionária, os $\hat{\boldsymbol{\rho}}_{\boldsymbol{ij}}(\boldsymbol{l})$ são estimativas amostrais dos valores teóricos de $\boldsymbol{\rho}_{\boldsymbol{ij}}(\boldsymbol{l})$; que são particularmente úteis na especificação do modelo para um processo vectorial média móvel de baixa ordem, uma vez que um processo MA(q) tem a propriedade de que $\boldsymbol{\rho}_{\boldsymbol{ij}}(\boldsymbol{l}) = \mathbf{0}$ para $l > q$.

Em termos de propriedades assimptóticas de correlações amostrais, salientas o seguinte caso que é estrema importância na análise dos resíduos: suponha-se que $\boldsymbol{Z_t}$ é um processo vectorial ruído branco, com matriz de covariância Σ e matriz de correlação $\rho(0)$, tal que $\boldsymbol{\rho_{ij}(l) = 0}$ para $l \neq 0$, então tem-se (Reinsel, 1997)

$$\boldsymbol{Var\left(\hat{\rho}_{ij}(l)\right) \approx \frac{1}{T}}$$

6.1.3.2 Matrizes de autoregressão parcial amostral e propriedades

Dado uma amostra de uma série temporal multivariada $\mathbf{Z_1}, \mathbf{Z_2}, \ldots, \mathbf{Z_T}$ de dimensão T, as estimativas das matrizes de autoregressão parcial serão bastante úteis na especificação da ordem de um modelo AR, uma vez que as matrizes $\boldsymbol{\Phi_{mm}}$ tem a propriedade de serem nulas para $l > p$ num processo AR(*p*). Para uma dada *lag m*, seja $\hat{\boldsymbol{\Phi}}_{\boldsymbol{mm}}$ a estimativa natural de $\boldsymbol{\Phi_{mm}}$ definido na secção 6.1.2.5, por exemplo por (6.21). Estas estimativas podem ser obtidas de forma óbvia substituindo os valores teóricos das matrizes de covariância $\boldsymbol{\Gamma(l)}$ pelas respectivas matrizes de covariância amostral $\hat{\boldsymbol{\Gamma}}(\boldsymbol{l})$ definidas na secção anterior.

Teste para a ordem de um modelo AR

No pressuposto de que o processo $\boldsymbol{Z_t}$ é um AR(p), então, para qualquer $m > p$, os elementos de $\hat{\boldsymbol{\Phi}}_{\boldsymbol{mm}}$ tem uma distribuição assimptoticamente normal com média zero, tal que $\hat{\boldsymbol{\phi}}_{mm} = vec\left(\hat{\boldsymbol{\Phi}}'_{mm}\right)$ tem matriz de covariância aproximada dada por $\boldsymbol{N^{-1}\left(\Sigma \otimes \Sigma^{*-1}_{m-1}\right)}$ onde $N = T - m$ e

$$\boldsymbol{\Sigma^{*}_{m-1} = \Gamma(0) - \Gamma^{*'}_{(m-1)}\Gamma^{-1}_{m-1}\Gamma^{*}_{(m-1)}}$$

Onde $\boldsymbol{\Gamma^{*}_{(m-1)} = \left(\Gamma(m-1), \ldots, \Gamma(1)\right)}$.. Consequentemente, substituindo $\boldsymbol{\Sigma^{*}_{m-1}}$ pela respectiva estimativa amostral, a estatística de *Wald* dada por

$$\begin{aligned} \boldsymbol{N\hat{\phi}'_{mm}\left(\hat{\Sigma}^{-1} \otimes \hat{\Sigma}^{*}_{m-1}\right)\hat{\phi}}_{mm} &= \boldsymbol{Ntr\left\{\hat{\Phi}_{mm}\hat{\Sigma}^{*}_{m-1}\hat{\Phi}'_{mm}\hat{\Sigma}^{-1}\right\}} \\ &\approx \boldsymbol{Ntr\left\{\hat{\Phi}_{mm}\hat{\Sigma}^{*}_{m-1}\hat{\Phi}'_{mm}\hat{\Sigma}^{-1}_{m-1}\right\}} \end{aligned}$$

Segue assimptoticamente uma distribuição χ^2 com k^2 graus de liberdade para $m > p$ na codição de um modelo AR(p) (Reinsel, 1997). Esta estatística pode ser utilizada como estatística de teste para testar um modelo AR(p), ou seja, para testar se $\boldsymbol{\Phi_m = 0}$ num modelo de ordem $m > p$.

Um teste equivalente é o chamado LR teste (*likelihood ratio test*) para $\boldsymbol{H_0}$: $\boldsymbol{\Phi_m = 0}$. Esta estatística LR é formulada em termos da razão $U_m = det(S_m)/det(S_{m-1})$ como $-N\log(U_m)$, com uma versão mais precisa para amostras finitas dada por

$$\boldsymbol{M_m = -[N - mk - 1 - 1/2]\log(U_m), \qquad N = T - m}$$

Onde $S_m = \sum_{t=m+1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}'_t$.

6.1.3.3 Estimativa dos mínimos quadrados condicional de modelos autoregressivos

O modelo AR(m) geral estacionário pode ser expresso na forma

$$(\boldsymbol{Z_t} - \boldsymbol{\mu}) = \sum_{\boldsymbol{j=1}}^{\boldsymbol{m}} \boldsymbol{\Phi_j}\big(\boldsymbol{Z_{t-j}} - \boldsymbol{\mu}\big) + \boldsymbol{\varepsilon_t} = \boldsymbol{\Phi}_{\boldsymbol{(m)}}^{'}\widetilde{\mathbf{X}}_{\mathbf{t}} + \boldsymbol{\varepsilon_t}$$

Onde se definiu $\widetilde{\mathbf{X}}_{\mathbf{t}} = [(\boldsymbol{Z_{t-1}} - \boldsymbol{\mu})^{'},\ \ldots\ ,\ (\boldsymbol{Z_{t-m}} - \boldsymbol{\mu})']^{'}$ e $\boldsymbol{\Phi}_{\boldsymbol{(m)}}^{'} = (\boldsymbol{\Phi_1},\ \ldots\ ,\boldsymbol{\Phi_m})$.

De forma equivalente, poder-se-ia escrever o modelo na forma

$$\boldsymbol{Z_t} = \boldsymbol{\delta} + \sum_{\boldsymbol{j=1}}^{\boldsymbol{m}} \boldsymbol{\Phi_j Z_{t-j}} + \boldsymbol{\varepsilon_t} = \boldsymbol{B'}\mathbf{X_t} + \boldsymbol{\varepsilon_t}$$

Com $\mathbf{X_t} = [\mathbf{1},\ \mathbf{Z}_{\mathbf{t-1}}^{'},\ \ldots\ ,\ \mathbf{Z}_{\mathbf{t-m}}^{'}]^{'}$ e $\boldsymbol{B}^{'} = (\boldsymbol{\delta}, \boldsymbol{\Phi_1},\ \ldots\ ,\boldsymbol{\Phi_m})$.

Definindo-se

1. Uma matriz de dados $\mathbf{Z} = [\mathbf{Z_{m+1}},\ \mathbf{Z_{m+2}}, \ldots\ , \mathbf{Z_T}]'$ de dimensão $N \times k$ com $N = T - m$
2. $\boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon_{m+1}},\ \boldsymbol{\varepsilon_{m+2}}, \ldots\ , \boldsymbol{\varepsilon_T}]'$
3. Uma matriz $\mathbf{X}$ de dimensão $N \times (mk + 1)$ cujas linhas típicas são $\boldsymbol{Y}_{\boldsymbol{t,m}}^{'} = [\mathbf{1},\ \mathbf{Z}_{\mathbf{t-1}}^{'},\ \ldots\ ,\ \mathbf{Z}_{\mathbf{t-m}}^{'}],\ t = m+1, \ldots\ , T$

Então tem-se $\mathbf{Z} = \mathbf{X}\boldsymbol{B} + \boldsymbol{\varepsilon}$, que tem a forma geral de um modelo linear multivariado com $N = T - m$ observações.

Segue-se que o estimador dos mínimos quadrados (*LS – Least Square, ou LSE - Least Square Estimator*) dos parâmetros AR são dados por

$$\boldsymbol{B} = \widehat{\boldsymbol{\Phi}}_{\mathbf{(m)}} = \big(\widehat{\boldsymbol{\Phi}}_{\mathbf{1}},\ \ \ \ \ldots\ , \widehat{\boldsymbol{\Phi}}_{\boldsymbol{m}}\big)^{'} = \big(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\big)^{-\mathbf{1}}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Z}} \qquad \textbf{(6.24)}$$

onde as matrizes $\widetilde{\boldsymbol{X}}$ e $\widetilde{\boldsymbol{Z}}$, respectivamente, tem linhas típicas $\big(Z_t - \bar{Z}_{(0)}\big)'$ e

$$\Big[\big(Z_t - \bar{Z}_{(1)}\big)^{'}, \ldots\ , \big(Z_{t-m} - \bar{Z}_{(m)}\big)^{'}\Big] \qquad t = m+1, \ldots\ , T$$

com $\bar{Z}_{(i)} = N^{-1}\sum_{t=m+1}^{T} Z_{t-i}$ e $n = M - m$. A estimativa de Σ é dada por

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{m}} = [N - (km+1)]^{-\mathbf{1}}\boldsymbol{S_m}$$

onde $\boldsymbol{S_m} = \sum_{t=m+1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t^{'}$ é a matriz da soma dos quadrados dos resíduos, com

$$\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{t}} = \big(\boldsymbol{Z_t} - \overline{\boldsymbol{Z}}_{\mathbf{(0)}}\big) - \sum_{\mathbf{j=1}}^{\boldsymbol{m}} \widehat{\boldsymbol{\Phi}}_{\mathbf{j}}\big(\boldsymbol{Z_{t-j}} - \overline{\boldsymbol{Z}}_{\boldsymbol{j}}\big)$$

o vector dos resíduos.

Este estimador $\widehat{\mathbf{\Phi}}_{\mathbf{j}}$ dos mínimos quadrados é também um estimador de máxima verosimilhança (ML – *Maximum Likelihood*), no pressuposto de uma distribuição normal. Para um modelo AR de ordem genérica, o estimador dos mínimos quadrados (LSE) $\widehat{\mathbf{\Phi}}$ tem aproximadamente tem aproximadamente as mesmas propriedades distributivas que o estimador dos mínimos quadrados para um modelo linear multivariado normal. Consequentemente, para um modelo estacionário AR(*m*), a distribuição de $vec\left(\widehat{\mathbf{\Phi}}_{(\mathbf{m})}\right)$ é aproximadamente normal multivariada com vector média $\mathbf{\Phi}_{(\mathbf{m})}$ e matriz de covariância que é consistentemente estimada por $\widehat{\mathbf{\Sigma}}_{\mathbf{m}} \otimes \left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\right)^{-\mathbf{1}}$. Ou seja, a inferência sobre os elementos individuais de $\boldsymbol{\phi} = \boldsymbol{vec}\left(\mathbf{\Phi}_{(\mathbf{m})}\right)$ pode ser baseada na distribuição aproximada

$$\widehat{\boldsymbol{\phi}} \sim \boldsymbol{N}\left(\boldsymbol{\phi}, \widehat{\mathbf{\Sigma}} \otimes \left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\right)^{-\mathbf{1}}\right)$$

em que $\widehat{\boldsymbol{\phi}} = \boldsymbol{vec}\left(\widehat{\mathbf{\Phi}}_{(\mathbf{m})}\right)$ (Reinsel, 1997).

A determinação da ordem p apropriada do modelo autoregressivo (AR) pode ser baseada na utilização aproximada do teste LR (*Likelihood Ratio Test*) descrito atrás. Por exemplo, quando se pretende testar a significância da matriz AR de *m*-ésima ordem, a hipótese nula é $H_0: \Phi_{\mathrm{m}} = 0$ contra a hipótese alternativa $H_0: \Phi_m \neq 0$. A estatística de teste será dada pela estatística LR

$$\boldsymbol{M_m} = -[\boldsymbol{N} - \boldsymbol{mk} - \mathbf{1} - \mathbf{1/2}] \log\left(\frac{|\boldsymbol{S_m}|}{|\boldsymbol{S_{m-1}}|}\right), \qquad \boldsymbol{N} = \boldsymbol{T} - \boldsymbol{m} \tag{6.25}$$

onde $\boldsymbol{S_{m-1}}$ é a matriz de soma dos quadrados dos resíduos obtida do ajustamento de um modelo AR de ordem $m-1$ para o mesmo conjunto de N observações utilizadas no ajustamento do modelo AR(m) (Box, Jenkins, & Reinsel, 2008) (Reinsel, 1997). Para valores elevados de N, quando $\Phi_{\mathrm{m}} = 0$ é verdadeiro, tem-se que $\boldsymbol{M_m}$ segue aproximadamente uma distribuição $\boldsymbol{\chi^2_{k^2}}$, e rejeita-se a hipótese nula para valores elevados de $\boldsymbol{M_m}$.

Consequentemente, o procedimento consiste em ajustar sucessivamente modelos AR de ordens superiores à série, e para cada ordem, testar sucessivamente a significância da última matriz coeficiente $\mathbf{\Phi}_{\mathbf{m}}$ incluída no modelo.

6.1.3.4 Técnicas adicionais para especificação de modelos VARMA

Após o ajustamento a um modelo AR dee cada uma das ordens $m = 1, 2, \ldots$, pelos mínimos quadrados, pode-se (deve-se) avaliar a qualidade do ajustamento de um modelo AR(m) considerando a sequência da estatística LM. Contudo este diagnóstico deve ser complementado pela análise dos resíduos, nomeadamente considerando os elementos da diagonal da matriz de covariância do ruído branco estimado $\hat{\Sigma}_{\mathrm{m}}$, que dá a indicação de como o aumento da ordem pode melhorar o ajustamento, uma vez que $\hat{\Sigma}_{\mathrm{m}}$ corresponde à variância do erro de previsão um passo à frente . Complementarmente, após cada ajustamento AR(*m*), $m = 1, 2, \ldots, M$, pode-se (deve-se) examinar as matrizes de correlação cruzada dos resíduos $\boldsymbol{\varepsilon_t}$ que darão informação adicional sobre a qualidade do ajustamento.

O processo $\{\boldsymbol{\varepsilon_t}\}$ Estas matrizes deverão ter um comportamento semelhante a um processo ruído branco.

De notar que para qualquer processo misto VARMA, pode ter que se recorrer a ajustamento autoregressivos de ordem bastante elevada antes de se conseguir um modelo que se pareça apropriado. Contudo, apesar dos actuais recursos computacionais, não serão desejáveis modelos finais de ordem muito elevada. Nesses casos, deve-se inspeccionar o modelo de correlação dos resíduos após ajustamentos AR de baixa ordem, com a possibilidade de detectar modelos mistos ARMA de baixa ordem que se verifiquem apropriados e simultaneamente reduzam o numero de parâmetros relativamente a um modelo AR puro.

Critérios de selecção de ordem para especificação de modelos
Das abordagens estudadas pode-se concluir que o teste LR associado à análise da correlação canónica parcial poderão ser bastante úteis para a determinação global da ordem AR nos casos em que um modelo de baixa ordem se verifique apropriado ao processo. Em situações mais complicadas, deve-se considerar modelos mistos ARMA que deverão posteriormente ser estimados recorrendo à estimativa de máxima verosimilhança. Há contudo a necessidade da determinação preliminar de uma adequada ordem (baixa) do modelo misto ARMA que se adeqúe às séries. Tal como no caso univariado, existem vários critérios, tais como AIC,BIC e FPE que normalmente podem ser úteis para a determinação das estruras de modelos mais apropriadas.

O critério AIC (normalizado po T) é dado por (Lütkepohl, 2007), (Reinsel, 1997)

$$\begin{aligned} \boldsymbol{AIC_{p,q}} &= \frac{\boldsymbol{-2\log(verosimilhança\ maximizada)+2r}}{\boldsymbol{T}} \\ &\approx \boldsymbol{\log\left(\left|\tilde{\Sigma}_r\right|\right)+\frac{2r}{T}+constante} \end{aligned} \quad (6.26)$$

onde r é o numero de parâmetros estimados por máxima verosimilhança para um modelo VARMA, e $\tilde{\Sigma}_r$ é a correspondente matriz de covariância estimada dos resíduos.

O critério BIC tem uma forma similar

$$\boldsymbol{BIC_r=\log\left(\left|\tilde{\Sigma}_r\right|\right)+r\frac{\log(T)}{T}} \quad (6.27)$$

O critério BIC penaliza mais o numero de parâmetros utilizado no modelo que o critério AIC. Um critério similar que está entre os dois anteriores é dado por

$$\boldsymbol{HQ_r=\log\left(\left|\tilde{\Sigma}_r\right|\right)+2r\frac{\log(\log(T))}{T}} \quad (6.28)$$

O critério FPE ("*Final Prediction Error*") para selecção do modelo VAR(*m*) é dado por

$$\boldsymbol{FPE_m=\det\left\{\left[1+\frac{mk}{T}\right]\hat{\Sigma}_m\right\}} \quad (6.29)$$

onde $\hat{\Sigma}_m=(\boldsymbol{T}/(\boldsymbol{T}-\boldsymbol{mk}))/\tilde{\Sigma}_r$ é a estimativa de Σ ajustada para os graus de liberade.

Estes critérios de selecção de modelo são utilizados para comparar vários modelos ajustados por máxima verosimilhança à série de dados, de modo a que se escolha o modelo que apresente valores mínimos para cada critério (Reinsel, 1997).

6.1.3.5 Modelos de regressão linear multivariada

Considere-se um sistema multivariado com *r* entradas e *k* saídas. Seja $\boldsymbol{Y_t} = (Y_{1t},\ Y_{2t},\ \dots\ , Y_{kt})'$ um vector aleatório de dimensão *k* das variáveis das respostas no instante *t*, e seja $\boldsymbol{X_t} = (X_{1t},\ X_{2t},\ \dots\ , X_{rt})'$ um vector de dimensão *r* das variáveis de entrada para o mesmo instante *t*. Considere-se um modelo linear multivariado da forma $Y_{it} = \boldsymbol{X_t'}\beta_{\mathrm{i}} + \varepsilon_{\mathrm{it}}$, com $i = 1, \dots, k$, ou, na notação matricial

$$\boldsymbol{Y_t'} = \boldsymbol{X_t'B} + \boldsymbol{\varepsilon_t'}, \qquad \boldsymbol{t} = \mathbf{1}, \dots\ , \boldsymbol{T}$$

Onde $\boldsymbol{B} = (\beta_1, \dots\ , \beta_{\mathrm{k}})$, e o $\boldsymbol{\varepsilon_t} = (\varepsilon_{1\mathrm{t}}, \dots\ , \varepsilon_{\mathrm{kt}})'$ são independentes e tem distribuição normal $N(0, \Sigma)$, tal que $\Sigma = E(\boldsymbol{\varepsilon_t}\boldsymbol{\varepsilon_t'})$. Dadas T observações $\boldsymbol{Y_1}, \dots\ , \boldsymbol{Y_T}$ e $\boldsymbol{X_1}, \dots\ , \boldsymbol{X_T}$, define-se a matriz de dados $\mathbf{Y} = (\boldsymbol{Y_1}, \dots\ , \boldsymbol{Y_T})'$ de dimensão $T \times k$, a matriz $\mathbf{X} = (\boldsymbol{X_1}, \dots\ , \boldsymbol{X_T})'$ de dimensão $T \times r$, e a matriz $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon_1}, \dots\ , \boldsymbol{\varepsilon_T})'$ de dimensão $T \times k$. Neste caso, o estimador de máxima verosimilhança (*MLE – Maximum Lilekihood Estimator*) de $\boldsymbol{B}$ é o mesmo que o estimador dos mínimos quadrados (LS), e que é dado por

$$\widehat{\boldsymbol{B}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y} \qquad \textbf{(6.30)}$$

Ou seja

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y_i} \qquad \boldsymbol{i} = \mathbf{1}, \dots\ , \boldsymbol{k}$$

Onde $\boldsymbol{y_i} = (Y_{i1},\ \dots\ , Y_{iT})'$ é a i-ésima coluna de $\mathbf{Y}$. O estimador para a matriz de covariância do erro Σ é dado por

$$\widehat{\boldsymbol{\Sigma}} = \frac{\mathbf{1}}{\boldsymbol{T-r}}\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{B}}\right)'\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{B}}\right) = \frac{\mathbf{1}}{\boldsymbol{T-r}}\sum_{\boldsymbol{t=1}}^{\boldsymbol{T}}\boldsymbol{\varepsilon_t}\boldsymbol{\varepsilon_t'}$$

De notar que a função de verosimilhança é dada por

$$\boldsymbol{L(B, \Sigma; Y)} = \frac{\mathbf{1}}{\boldsymbol{(2\pi)^{kT/2}|\Sigma|^{T/2}}}\boldsymbol{exp}\left[-\frac{\mathbf{1}}{\mathbf{2}}\sum_{\boldsymbol{t=1}}^{\boldsymbol{T}}\boldsymbol{(Y_t - B'X_t)'\Sigma^{-1}(Y_t - B'X_t)}\right] \qquad \textbf{(6.31)}$$

Que será maximizada para $\boldsymbol{B}$ igual a $\widehat{\boldsymbol{B}}$ em (6.30) e $\boldsymbol{\Sigma}$ igual a $\widetilde{\boldsymbol{\Sigma}} = [(\boldsymbol{T-r})/\boldsymbol{T}]\widehat{\boldsymbol{\Sigma}}$ (Reinsel, 1997)**.**

Um modo bastante conveniente de lidar com regressão linear multivariada principalmente quando se impõem restrições nos parâmetros, é transformar o modelo linear multivariado na forma de modelo univariado. Partindo da matriz $T \times k$ $\mathbf{Y} = (\boldsymbol{y_1}, \dots\ , \boldsymbol{y_k})$ definida acima, define-se $\boldsymbol{y} = \boldsymbol{vec}(\mathbf{Y}) = (\boldsymbol{y_1'}, \dots\ , \boldsymbol{y_k'})'$ e $\boldsymbol{e} = \boldsymbol{vec}(\boldsymbol{\varepsilon}) = (\boldsymbol{e_1'}, \dots\ , \boldsymbol{e_k'})'$ onde $\boldsymbol{e_i}$ é a *i*-ésima coluna de $\boldsymbol{\varepsilon}$. Relembrando as propriedades relacionadas com o operador "*vec*", tem-se que se $Z = ABC$, então $\boldsymbol{vec(Z)} = \boldsymbol{vec(ABC)} = (\boldsymbol{C'} \otimes \boldsymbol{A})\boldsymbol{vec(B)}$, e também que $\boldsymbol{tr(ABCB')} = [\boldsymbol{vec(B')}]'(\boldsymbol{C'} \otimes \boldsymbol{A})\boldsymbol{vec(B)}$.

Da expressão linear $\mathbf{Y} = \mathbf{X}\boldsymbol{B} + \boldsymbol{\varepsilon}$, tem-se que

$$\mathbf{y} = \boldsymbol{vec}(\mathbf{Y}) = \boldsymbol{vec}(\mathbf{X}\boldsymbol{B}) + \boldsymbol{vec}(\boldsymbol{\varepsilon}) = (\boldsymbol{I_k} \otimes \mathbf{X})\boldsymbol{\beta} + \mathbf{e}$$

onde obviamente $\boldsymbol{\beta} = \boldsymbol{vec}(\boldsymbol{B})$. De notar que $\boldsymbol{Cov}(\mathbf{e}) = \boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \boldsymbol{I_T}$ ou seja $\boldsymbol{Cov}(\mathbf{e_i}, \mathbf{e_j}) = \boldsymbol{\sigma_{ij}}\boldsymbol{I_T}$, onde $\boldsymbol{\sigma_{ij}}$ é o elemento (*i,j*) de $\boldsymbol{\Sigma}$. O modelo assim definido tem a forma do modelo de regressão linear univariado geral, então o estimador de máxima verosimilhança de $\boldsymbol{\beta}$ (MLE) é o mesmo que o estimador dos mínimos quadrados generalizado (Reinsel, 1997)

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \left[(\boldsymbol{I_k} \otimes \mathbf{X})'\boldsymbol{\Omega^{-1}}(\boldsymbol{I_k} \otimes \mathbf{X})\right]^{-1}(\boldsymbol{I_k} \otimes \mathbf{X})'\boldsymbol{\Omega^{-1}}\mathbf{y} \\ &= \left(\boldsymbol{I_k} \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\boldsymbol{y}\end{aligned}$$

Tem-se então que $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_{\boldsymbol{i}},\ \boldsymbol{i} = \mathbf{1}, \ldots, \boldsymbol{k}$ é um estimador dos mínimos quadrados para cada *i*. a estimativa de $\boldsymbol{\beta}$ obtida por este método tem distribuição normal multivariada com média

$$\boldsymbol{E}\left(\widehat{\boldsymbol{\beta}}\right) = \left(\boldsymbol{I_k} \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{E}(\mathbf{y}) = \boldsymbol{\beta}$$

e covariância

$$\begin{aligned}\boldsymbol{Cov}\left(\widehat{\boldsymbol{\beta}}\right) &= \left(\boldsymbol{I_k} \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)(\boldsymbol{\Sigma} \otimes \boldsymbol{I_T})\left(\boldsymbol{I_k} \otimes \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right) \\ &= \left(\boldsymbol{\Sigma} \otimes (\mathbf{X}'\mathbf{X})^{-1}\right)\end{aligned}$$

ou seja, $\boldsymbol{E}\left(\widehat{\boldsymbol{\beta}}_{\mathbf{i}}\right) = \boldsymbol{\beta}_{\mathbf{i}}$ e $\boldsymbol{Cov}\left(\widehat{\boldsymbol{\beta}}_{\mathbf{i}}, \widehat{\boldsymbol{\beta}}_{\mathbf{j}}\right) = \boldsymbol{\sigma_{ij}}(\mathbf{X}'\mathbf{X})^{-1}$. Como a matriz de covariância, na prática, é desconhecida, a matriz de covariância de $\widehat{\boldsymbol{\beta}}$ é estimada por $\widehat{\boldsymbol{Cov}}\left(\widehat{\boldsymbol{\beta}}\right) = \left(\widehat{\boldsymbol{\Sigma}} \otimes (\mathbf{X}'\mathbf{X})^{-1}\right)$ (Reinsel, 1997).

Restrições lineares

Na estimação de parâmetros, normalmente torna-se necessário restringir alguns parâmetros, ou porque o valor deles é conhecido, por exemplo em modelos mistos determinístico/estocástico, ou porque os seus valores não são significativos, e portanto serão nulos. Consequentemente, aos estimar-se os valores dos parâmetros do modelo a ajustar, essas restrições tem de ser impostas.

Considere-se o modelo linear, seguindo a notação de (Lütkepohl, 2007)

$$\boldsymbol{y_t} = \boldsymbol{\nu} + \boldsymbol{A_1 y_{t-1}}, \ldots, \boldsymbol{A_p y_{t-p}} + \boldsymbol{u_t}$$

que pode ser escrito na forma compacta

$$\boldsymbol{Y} = \boldsymbol{BZ} + \boldsymbol{U} \qquad \textbf{(6.32)}$$

onde

$$\boldsymbol{Y} = [\boldsymbol{y_1}, \dots, \boldsymbol{y_T}]$$
$$\boldsymbol{Z} = [\boldsymbol{Z_0}, \dots, \boldsymbol{y_{T-1}}]$$
$$\boldsymbol{B} = [\boldsymbol{\nu}, \boldsymbol{A_1}, \dots, \boldsymbol{A_p}]$$
$$\boldsymbol{U} = [\boldsymbol{u_1}, \dots, \boldsymbol{u_T}]$$

Em que cada coluna de é composta por

$$\boldsymbol{Z_t} = \begin{bmatrix} \mathbf{1} \\ \boldsymbol{y_t} \\ \vdots \\ \boldsymbol{y_{t-p+1}} \end{bmatrix}$$

Define-se então uma restrição linear para $\boldsymbol{B}$ na forma

$$\boldsymbol{\beta} = \boldsymbol{vec}\,(\boldsymbol{B}) = \boldsymbol{R\gamma} + \boldsymbol{r} \qquad \textbf{(6.33)}$$

Onde $\boldsymbol{\beta} = \boldsymbol{vec}\,(\boldsymbol{B})$ é um vector de dimensão $K(Kp+1) \times 1$, R é uma matriz conhecida de dimensão $K(Kp+1) \times M$ de rank M, γ é um vector não restrito de dimensão $M \times 1$ dos parâmetros desconhecidos e r é um vector de dimensão $K(Kp+1) \times 1$ de constantes conhecidas. Todas as restrições lineares de interesse podem ser escritas nesta forma. Por exemplo, a restrição $A_p = 0$ pode ser escrita na forma (6.33) escolhendo $M = K^2(p-1) + k$,

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{I_M} \\ \mathbf{0} \end{bmatrix}, \quad \boldsymbol{\gamma} = \boldsymbol{vec}(\boldsymbol{\nu}, \boldsymbol{A_1}, \dots, \boldsymbol{A_{p-1}})$$

e $r = 0$. Mais genericamente, concluindo-se que o *i*-ésimo elemento do vector $\boldsymbol{\beta}$ é não significativo, e portanto nulo, para construir esta restrição, procede-se do seguinte modo:

1. Define-se a matriz $\boldsymbol{R}$ a partir de uma matriz identidade $\left(\boldsymbol{I_{K(Kp+1)}}\right)$ de dimensão $K(Kp+1) \times K(Kp+1)$, em que a *i*-ésima coluna foi retirada,
2. Define-se o vector $\boldsymbol{\gamma}$ a partir do vector $\boldsymbol{vec}\,(\boldsymbol{B})$ em que o *i*-ésimo elemento foi retirado
3. Define-se $r = 0$.

A representação (6.33) permite impor as restrições pela simples reparametrização do modelo original. Aplicando o operador $\boldsymbol{vec}$ a (6.32) e substituindo $\boldsymbol{\beta}$ por $\boldsymbol{R\gamma} + \boldsymbol{r}$ tem-se

$$\begin{aligned} \mathbf{y} &= \boldsymbol{vec}(\boldsymbol{Y}) = (\mathbf{Z}' \otimes \boldsymbol{I_K})\boldsymbol{vec}(\boldsymbol{B}) + \boldsymbol{vec}(\boldsymbol{U}) \\ &= (\mathbf{Z}' \otimes \boldsymbol{I_K})(\boldsymbol{R\gamma} + \boldsymbol{r}) + \mathbf{u} \end{aligned}$$

ou

$$\boldsymbol{z} = (\mathbf{Z}' \otimes \boldsymbol{I_K})\boldsymbol{R\gamma} + \mathbf{u} \qquad \textbf{(6.34)}$$

onde $\mathbf{z} = \mathbf{y} - (\mathbf{Z}' \otimes \boldsymbol{I_K})\boldsymbol{r}$ e $\mathbf{u} = \boldsymbol{vec}(\boldsymbol{U})$. Esta forma de modelo permite derivar permite derivar estimadores e respectivas propriedades tal como nos modelos originais não restitos (Lütkepohl, 2007).

Estimativa GLS (*generalized LS*) e EGLS (*estimated GLS*)

Denotando por $\boldsymbol{\Sigma_u}$ a matriz de covariância de $\boldsymbol{u_t}$, o vector de parâmetros $\hat{\boldsymbol{\gamma}}$ que minimiza

$$\begin{aligned} \boldsymbol{S(\gamma)} &= \boldsymbol{u'}\left(\boldsymbol{I_T} \otimes \boldsymbol{\Sigma_u^{-1}}\right)\boldsymbol{u} \\ &= [\boldsymbol{z} - (\boldsymbol{Z'} \otimes \boldsymbol{I_K})\boldsymbol{R\gamma}]'\left(\boldsymbol{I_T} \otimes \boldsymbol{\Sigma_u^{-1}}\right)[\boldsymbol{z} - (\boldsymbol{Z'} \otimes \boldsymbol{I_K})\boldsymbol{R\gamma}] \end{aligned} \quad (6.35)$$

com respeito a γ é dado por (Lütkepohl, 2007)

$$\begin{aligned} \hat{\boldsymbol{\gamma}} &= \left[\boldsymbol{R'}\left(\boldsymbol{ZZ'} \otimes \boldsymbol{\Sigma_u^{-1}}\boldsymbol{R}\right)\right]^{-1}\boldsymbol{R'}\left(\boldsymbol{Z'} \otimes \boldsymbol{\Sigma_u^{-1}}\right)\boldsymbol{z} \\ &= \boldsymbol{\gamma} + \left[\boldsymbol{R'}\left(\boldsymbol{ZZ'} \otimes \boldsymbol{\Sigma_u^{-1}}\boldsymbol{R}\right)\right]^{-1}\boldsymbol{R'}\left(\boldsymbol{I_{Kp+1}} \otimes \boldsymbol{\Sigma_u^{-1}}\right)\boldsymbol{vec}(\boldsymbol{UZ'}) \end{aligned} \quad (6.36)$$

Este estimador é conhecido como estimador dos mínimos quadrados generalizados (GLS - *Generalized Seast Square*) porque minimiza a soma dos erros quadrados generalizados em vez da soma dos erros quadrados $\boldsymbol{u'u}$, e geralmente é assimptoticamente mais que o estimador dos mínimos quadrados multivariado. Sob o pressuposto da normalidade dos dados o estimador GLS é equivalente ao estimador da máxima verosimilhança.

Segundo (Lütkepohl, 2007, p. 196), nas condições de que $\boldsymbol{y_t}$ é um processo de dimensão K estável VAR(*p*), estacionário, e $\boldsymbol{u_t}$ é ruído branco independente com quartos momentos limitados. Se $\boldsymbol{\beta} = \boldsymbol{vec}\,(\boldsymbol{B}) = \boldsymbol{R\gamma} + \boldsymbol{r}$ como em (6.33) com rank de R igual a M, então o vector $\hat{\gamma}$ dada por **(6.36**) é um estimador consistente de $\boldsymbol{\gamma}$ e

$$\sqrt{\boldsymbol{T}}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \xrightarrow{\boldsymbol{d}} \boldsymbol{N}\left(\boldsymbol{0}, \left[\boldsymbol{R'}\left(\boldsymbol{\Gamma} \otimes \boldsymbol{\Sigma_u^{-1}}\right)\boldsymbol{R}\right]^{-1}\right) \quad \textbf{(6.37)}$$

onde $\xrightarrow{\boldsymbol{d}}$ significa que "converge em distribuição para" e $\boldsymbol{\Gamma} := \boldsymbol{E}(\boldsymbol{Z_t}\boldsymbol{Z_t'})$ é igual ao limite de probabilidade ZZ'/T.

A determinação do estimador $\hat{\gamma}$ requer o conhecimento de matriz de covariância $\boldsymbol{\Sigma_u}$ que normalmente é desconhecia, pelo terá que ser substituída por uma estimativa. Recorrendo-se a um estimador consistente $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}$ de para substituir $\boldsymbol{\Sigma_u}$ em (6.36) um estimador EGLS (*Estimated GLS*)

$$\hat{\hat{\boldsymbol{\gamma}}} = \left[\boldsymbol{R'}\left(\boldsymbol{ZZ'} \otimes \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\boldsymbol{R}\right)\right]^{-1}\boldsymbol{R'}\left(\boldsymbol{Z'} \otimes \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\right)\boldsymbol{z}$$

Que tem as mesmas propriedades assimptóticas que o estimador $\hat{\gamma}$

Um estimador consistente de $\boldsymbol{\Sigma_u}$ pode ser determinado através da expressão (Lütkepohl, 2007)

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{u}} &= \frac{\boldsymbol{1}}{\boldsymbol{T - Kp - 1}}\left(\boldsymbol{Y} - \widehat{\boldsymbol{B}}\boldsymbol{Z}\right)\left(\boldsymbol{Y} - \widehat{\boldsymbol{B}}\boldsymbol{Z}\right)' \\ &= \frac{\boldsymbol{1}}{\boldsymbol{T - Kp - 1}}\boldsymbol{Y}\left(\boldsymbol{I_T} - \boldsymbol{Z'}(\boldsymbol{Z'}\boldsymbol{Z})^{-1}\boldsymbol{Z}\right)\boldsymbol{Y'} \end{aligned} \quad (6.38)$$

Onde se utilizou a igualdade $\widehat{\boldsymbol{B}} = \boldsymbol{YZ}'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ que é o estimador dos mínimos quadrados (LS) multivariado não restrito das matrizes coeficiente $\boldsymbol{B}$.

Uma hipótese alternativa consiste em determinar num primeiro passo um estimador dos mínimos quadrados (LS) que minimize $\mathbf{u}'\mathbf{u}$ com respeito a $\boldsymbol{\gamma}$, que obviamanete será dado por

$$\widehat{\boldsymbol{\gamma}} = [\boldsymbol{R}'(\boldsymbol{Z}\boldsymbol{Z}' \otimes \boldsymbol{I}_{\boldsymbol{K}}\boldsymbol{R})]^{-1}\boldsymbol{R}'(\boldsymbol{Z}' \otimes \boldsymbol{I}_{\boldsymbol{K}})\boldsymbol{z}$$

6.1.4 Estimativa de máxima verosimilhança e validação do modelo

Como se verificou em secções anteriores, a estimação de máxima verosimilhança para os parâmetros de modelos ARMA, tanto univariados como multivariados conduz a equações não lineares. Nesta secção apresenta-se o procedimento para estimação de máxima verosimilhança condicional dos parâmetros para o modelo ARMA e respectivas propriedades. Examina-se a computação explicita da estimativa de máxima verosimilhança condicional através do desenvolvimento explicito de um procedimento Newton-Raphson (Gauss-Newton) modificado.

6.1.4.1 Função condicional de máxima verosimilhança para o modelo VARMA

Considere-se a estimação de parâmetros da máxima verosimilhança condicional para o modelo VARMA(*p,q*) (assumindo média zero por conveniência),

$$\boldsymbol{Z_t} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\Phi_j Z_{t-j}} = \boldsymbol{\varepsilon_t} - \sum_{\mathbf{j=1}}^{\mathbf{q}} \boldsymbol{\Theta_j}\, \boldsymbol{\varepsilon_{t-j}} \tag{6.39}$$

Baseada numa amostra de T observações vectoriais $\boldsymbol{Z_t}$, $t = 1, 2, \ldots, T$. A abordagem de máxima verosimilhança condicional baseia-se na suposição de que as observações iniciais $\boldsymbol{Z_0}, \boldsymbol{Z_{-1}}, \ldots, \boldsymbol{Z_{-p+1}}$ estão também disponíveis (por conveniência de notação) e são consideradas como fixas, utiliza-se ainda a aproximação para os distúrbios iniciais $\varepsilon_0 = \varepsilon_{-1} = \cdots = \varepsilon_{-q+1} = 0$, pelo que T será o numero efectivo de observações. Assume-se, como é costume, que os ε_t são independentes e normalmente distribuídos com média zero e matriz de covariância não singular $\boldsymbol{\Sigma}$.

Define-se as matrizes $T \times k$ $\mathbf{Z} = (\boldsymbol{Z_1}, \ldots, \boldsymbol{YZ_T})$ e $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon_1}, \ldots, \boldsymbol{\varepsilon_T})$, com a notação $\boldsymbol{\mathcal{B}^i}\mathbf{Z} = (\boldsymbol{Z_{1-i}}, \ldots, \boldsymbol{Z_{T-i}})$ e $\boldsymbol{\mathcal{B}^i\varepsilon} = (\boldsymbol{\varepsilon_{1-i}}, \ldots, \boldsymbol{\varepsilon_{T-i}})$. Deste modo, o modelo pode ser expresso na forma

$$\mathbf{Z} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \boldsymbol{\mathcal{B}^i Z \Phi_i'} = \boldsymbol{\varepsilon} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{q}} \boldsymbol{\mathcal{B}^i \varepsilon \Theta_i'} \tag{6.40}$$

Recorrendo novamente ao operador "*vec*" e à relação $\boldsymbol{vec(ABC)} = (\boldsymbol{C}' \otimes \boldsymbol{A})\boldsymbol{vec(B)}$, o modelo poderá ser expresso em forma de vector. Defina-se os vectores

$$\boldsymbol{z} = \boldsymbol{vec}(\mathbf{Z}) = (\boldsymbol{Z_1'}, \ldots, \boldsymbol{Z})'$$

$$\boldsymbol{e} = \boldsymbol{vec}(\boldsymbol{\varepsilon}) = (\boldsymbol{\varepsilon_1'}, \ \dots \ , \boldsymbol{\varepsilon_k'})'$$
$$\boldsymbol{\mathcal{B}^i}\mathbf{z} = \boldsymbol{vec}\left(\boldsymbol{\mathcal{B}^i}\mathbf{Z}\right)$$
$$\boldsymbol{\mathcal{B}^i}\mathbf{e} = \boldsymbol{vec}\left(\boldsymbol{\mathcal{B}^i}\boldsymbol{\varepsilon}\right)$$
$$\boldsymbol{\phi} = \boldsymbol{vec}(\boldsymbol{\Phi_i}) \qquad i = 1, \dots , p$$
$$\boldsymbol{\theta} = \boldsymbol{vec}(\boldsymbol{\Theta_i}) \qquad i = 1, \dots , q$$

O modelo pode então ser expresso na forma

$$\mathbf{z} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} (\boldsymbol{I_T} \otimes \boldsymbol{\Phi_i})\boldsymbol{\mathcal{B}^i}\mathbf{z} = \mathbf{e} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{q}} (\boldsymbol{I_T} \otimes \boldsymbol{\Theta_i})\,\boldsymbol{\mathcal{B}^i}\mathbf{e} \tag{6.41}$$

ou na forma

$$\mathbf{z} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \left(\boldsymbol{\mathcal{B}^i}\mathbf{Z} \otimes \boldsymbol{I_k}\right)\boldsymbol{\phi_i} = \mathbf{e} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{q}} \left(\boldsymbol{\mathcal{B}^i}\boldsymbol{\varepsilon} \otimes \boldsymbol{I_k}\right)\boldsymbol{\theta_i} \tag{6.42}$$

Introduz-se a matriz L de atraso ou desfasagem $T \times T$, que é composta por uns na sub-diagonal imediatamente a seguir à diagonal principal e zeros nos restantes sítios. Considerando os valores iniciais de $\boldsymbol{\varepsilon_t}$ iguais a zero, o parâmetro $\boldsymbol{\mathcal{B}^i}\boldsymbol{\varepsilon}$ do segundo membro de (6.42) pode ser substituído por $\boldsymbol{L^i}\boldsymbol{\varepsilon} = (\mathbf{0}, \ \dots \ , \mathbf{0}, \boldsymbol{\varepsilon_1}, \ \dots \ , \boldsymbol{\varepsilon_{T-i}})'$ e consequentemente, o termo $\mathcal{B}^i\mathrm{e}$ em (6.41) fica na forma $\left(\boldsymbol{L^i}\boldsymbol{\varepsilon} \otimes \boldsymbol{I_k}\right)\mathbf{e} = (\mathbf{0}', \ \dots \ , \mathbf{0}', \boldsymbol{\varepsilon_1'}, \ \dots \ , \boldsymbol{\varepsilon_{T-i}'})'$. Das formas (6.42) e (6.41) obtém-se a relação

$$\begin{aligned} \mathbf{z} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \left(\boldsymbol{\mathcal{B}^i}\mathbf{Z} \otimes \boldsymbol{I_k}\right)\boldsymbol{\phi_i} \ &= \mathbf{e} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{q}} \left(\boldsymbol{L^i}\boldsymbol{\varepsilon} \otimes \boldsymbol{I_k}\right)\boldsymbol{\theta_i} \\ &= \mathbf{e} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{q}} \left(\boldsymbol{L^i} \otimes \boldsymbol{\Theta_i}\right)\mathbf{e} = \boldsymbol{\Theta}\mathbf{e} \end{aligned} \tag{6.43}$$

Onde se definiu $\boldsymbol{\Theta} = (\boldsymbol{I_T} \otimes \boldsymbol{I_k}) - \sum_{\boldsymbol{i=1}}^{\boldsymbol{q}} \left(\boldsymbol{L^i} \otimes \boldsymbol{\Theta_i}\right)$. No pressuposto da normalidade dos resíduos ε_t, ou seja, $\mathbf{e} \sim \mathbf{N}\left(\mathbf{0}, \ L^i \otimes \boldsymbol{\Sigma}\right)$, a função de máxima verosimilhança (condicional) pode ser escrita na forma (Reinsel, 1997) (ver expressão (6.31), tomando-se o logaritmo)

$$\begin{aligned} \boldsymbol{l} \ &= -\frac{\boldsymbol{T}}{\mathbf{2}}\mathbf{log}|\boldsymbol{\Sigma}| - \frac{\mathbf{1}}{\mathbf{2}}\sum_{\boldsymbol{t=1}}^{\boldsymbol{T}} \boldsymbol{\varepsilon_t'}\boldsymbol{\Sigma^{-1}}\boldsymbol{\varepsilon_t} \\ &= -\frac{\boldsymbol{T}}{\mathbf{2}}\mathbf{log}|\boldsymbol{\Sigma}| - \frac{\mathbf{1}}{\mathbf{2}}\mathbf{e}'\left(\boldsymbol{I_T} \otimes \boldsymbol{\Sigma^{-1}}\right)\mathbf{e} \\ &= -\frac{\boldsymbol{T}}{\mathbf{2}}\mathbf{log}|\boldsymbol{\Sigma}| - \frac{\mathbf{1}}{\mathbf{2}}\mathbf{w}'\boldsymbol{\Theta}'^{-\mathbf{1}}\left(\boldsymbol{I_T} \otimes \boldsymbol{\Sigma^{-1}}\right)\mathbf{w}\boldsymbol{\Theta^{-1}} \end{aligned} \tag{6.44}$$

Onde, a partir de (6.43) se definiu $\mathrm{w} = \mathrm{z} - \sum_{i=1}^{p}\left(\mathcal{B}^i\mathbf{Z} \otimes I_k\right)\phi_i = \Theta\mathrm{e}$, com $\mathbf{w} = (\boldsymbol{W_1'}, \ \dots \ , \boldsymbol{W_T'})'$ e $\boldsymbol{W_t} = \boldsymbol{Z_t} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \boldsymbol{\Phi_i}\boldsymbol{Z_{t-i}}$.

Para se determinar as equações de máxima verosimilhança tem que se maximizar a função de máxima verosimilhança (6.44) com respeito aos parâmetros ϕ_j, θ_j e Σ. Para ϕ_j, θ_j fixos, a maximização com respeito a Σ é dada por

$$\widehat{\boldsymbol{\Sigma}} = \frac{\mathbf{1}}{\boldsymbol{T}}\sum_{\boldsymbol{t=1}}^{\boldsymbol{T}} \boldsymbol{\varepsilon_t}\boldsymbol{\varepsilon_t}' = \frac{\mathbf{1}}{\boldsymbol{T}}\varepsilon'\varepsilon$$

onde $\boldsymbol{vec}(\boldsymbol{\varepsilon}') = \mathbf{e} = \boldsymbol{\Theta}^{-\mathbf{1}}\mathbf{w}$. As derivadas parciais de l com respeito a ϕ_j e a θ_j são dadas por $(\partial \boldsymbol{l}/\partial\phi_j) = -(\partial\boldsymbol{\varepsilon}'/\partial\phi_j)(I_T \otimes \Sigma^{-1})\mathbf{e}$ e $(\partial \boldsymbol{l}/\partial\phi_j) = -(\partial\boldsymbol{\varepsilon}'/\partial\,\theta_j)(I_T \otimes \Sigma^{-1})\mathbf{e}$, pelo que, de (6.43) se tem (Reinsel, 1997)

$$\frac{\partial \boldsymbol{l}}{\partial \boldsymbol{\phi}_j} = (\boldsymbol{\mathcal{B}^j}\mathbf{Z} \otimes \boldsymbol{I_k})'\boldsymbol{\Theta}'^{-\mathbf{1}}(\boldsymbol{I_T} \otimes \boldsymbol{\Sigma^{-1}})\boldsymbol{\Theta^{-1}}\left(\mathbf{z} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}}(\boldsymbol{\mathcal{B}^i}\mathbf{Z} \otimes \boldsymbol{I_k})\boldsymbol{\phi}_i\right)$$

Para $j = 1, \ldots, p$, e

$$\frac{\partial \boldsymbol{l}}{\partial \theta_j} = -(\boldsymbol{L^j}\varepsilon \otimes \boldsymbol{I_k})'\boldsymbol{\Theta}'^{-\mathbf{1}}(\boldsymbol{I_T} \otimes \boldsymbol{\Sigma^{-1}})\boldsymbol{\Theta^{-1}}\left(\mathbf{z} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}}(\boldsymbol{\mathcal{B}^i}\mathbf{Z} \otimes \boldsymbol{I_k})\boldsymbol{\phi}_i\right)$$

Para $j = 1, \ldots, q$.

Definindo-se o vector $\boldsymbol{\beta}$ da forma

$$\boldsymbol{\beta} = (\boldsymbol{\phi_1'}, \ldots, \boldsymbol{\phi_p'}, \boldsymbol{\theta_1'}, \ldots, \boldsymbol{\theta_q'})'$$

E a matriz

$$\mathbf{Y} = [(\boldsymbol{\mathcal{B}}\mathbf{Z} \otimes \boldsymbol{I_k}), \ldots, (\boldsymbol{\mathcal{B}^p}\mathbf{Z} \otimes \boldsymbol{I_k}), -(L\boldsymbol{\varepsilon} \otimes \boldsymbol{I_k}), \ldots, -(\boldsymbol{L^q}\boldsymbol{\varepsilon} \otimes \boldsymbol{I_k})]$$

Deste modo, estas derivadas podem ser expressas conjuntamente, numa forma mais conveniente

$$\begin{aligned}\frac{\boldsymbol{\partial l}}{\boldsymbol{\partial \beta}} &= \mathbf{Y}'\boldsymbol{\Theta}'^{-\mathbf{1}}(\boldsymbol{I_T} \otimes \boldsymbol{\Sigma^{-1}})\boldsymbol{\Theta^{-1}}\left(\mathbf{z} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}}(\boldsymbol{\mathcal{B}^i}\mathbf{Z} \otimes \boldsymbol{I_k})\boldsymbol{\phi_i}\right) \\ &= \mathbf{Y}'\boldsymbol{\Theta}'^{-\mathbf{1}}(\boldsymbol{I_T} \otimes \boldsymbol{\Sigma^{-1}})\mathbf{e} \qquad (6.45) \\ &= -\sum_{\mathbf{t=1}}^{\mathbf{T}}\frac{\boldsymbol{\partial \varepsilon_t'}}{\boldsymbol{\partial \beta}}\boldsymbol{\Sigma^{-1}}\boldsymbol{\varepsilon_t}\end{aligned}$$

No caso de se ter q=0, ou seja, no caso de se ter um processo autoregressivo AR(p) puro com $\boldsymbol{\Theta} = \boldsymbol{I_{kT}}$, tem-se que $\boldsymbol{\beta} = \boldsymbol{vec}(\boldsymbol{\Phi}_{(\boldsymbol{p})}')$ na notação de 6.1.3 com $\boldsymbol{\Phi}_{(\boldsymbol{p})}' = (\boldsymbol{\Phi_1}, \ldots, \boldsymbol{\Phi_p})$ e $\mathbf{Y} = \mathbf{X} \otimes \boldsymbol{I_k}$, onde $\mathbf{X} = (\boldsymbol{\mathcal{B}}\mathbf{Z}, \ldots, \boldsymbol{\mathcal{B}^p}\mathbf{Z})$. Então a função de máxima verosimilhança simplifica-se ganhando a forma

$$\frac{\partial \boldsymbol{l}}{\partial \boldsymbol{\beta}} = \mathbf{Y}'\left(\boldsymbol{I_T} \otimes \boldsymbol{\Sigma^{-1}}\right)(\boldsymbol{z} - \mathbf{Y}\boldsymbol{\beta}) = \left(\boldsymbol{I_{kp}} \otimes \boldsymbol{\Sigma^{-1}}\right)\mathbf{Y}'(\boldsymbol{z} - \mathbf{Y}\boldsymbol{\beta}) = \mathbf{0}$$

Que conduz à estimativa de máxima verosimilhança condicional de um modelo AR(p) dada por

$$\widehat{\boldsymbol{\beta}} = (\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\boldsymbol{z} = \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \otimes \boldsymbol{I_k}\right)\mathbf{z}$$

Que é o mesmo estimador que o estimador dos mínimos quadrados.

Como já foi referido diversas vezes, inclusive na abordagem a sistemas univariados, para $q > 0$, as equações de máxima verosimilhança, neste caso as equações (6.45), são altamente não lineares nos parâmetros $\boldsymbol{\beta}$, pelo que tem de ser resolvidas com recurso a procedimentos numéricos iterativos tais como o método Newton-Raphson cujas equações para um estimador de máxima verosimilhança (aproximado) para $\widehat{\boldsymbol{\beta}}$ são (Reinsel, 1997)

$$-\left(\frac{\partial^2 \boldsymbol{l}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\right)_{\boldsymbol{\beta_0}} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}\right) = \left(\frac{\partial \boldsymbol{l}}{\partial \boldsymbol{\beta}}\right)_{\boldsymbol{\beta_0}} \tag{6.46}$$

onde $\boldsymbol{\beta_0}$ é uma estimativa inicial de $\boldsymbol{\beta}$ e a estimativa de $\boldsymbol{\Sigma}$, necessária em (6.45) pode-se obter a partir da iteração anterior por $\widetilde{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{\varepsilon}}'\tilde{\boldsymbol{\varepsilon}}$.

Determinação iterativa da estimativa condicional MLE

Para implementar o método iterativo *Newton-Raphson* (6.46), é seria útil encontrar uma expressão conveniente para a matriz Hessiana de segundas derivadas parciais. Desprezando os termos de menor ordem que dividem por T, uma vez que tendem para zero quando T tende para infinito, obtém-se uma aproximação para a Hessiana

$$-\left(\frac{\partial^2 \boldsymbol{l}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\right) \approx \left(\frac{\partial \mathbf{e}'}{\partial \boldsymbol{\beta}}\right)\left(\boldsymbol{I_T} \otimes \boldsymbol{\Sigma^{-1}}\right)\left(\frac{\partial \mathbf{e}}{\partial \boldsymbol{\beta}'}\right) = \sum_{\boldsymbol{t=1}}^{\boldsymbol{T}} \left(\frac{\partial \boldsymbol{\varepsilon}_{\boldsymbol{t}}'}{\partial \boldsymbol{\beta}}\right)\boldsymbol{\Sigma^{-1}}\left(\frac{\partial \boldsymbol{\varepsilon}_{\mathbf{t}}}{\partial \boldsymbol{\beta}'}\right)$$

$$= \mathbf{Y}'\boldsymbol{\Theta}'^{-1}\left(\boldsymbol{I_T} \otimes \boldsymbol{\Sigma^{-1}}\right)\boldsymbol{\Theta^{-1}}\mathbf{Y} \tag{6.47}$$

Esta aproximação implica a negligência de termos que tem a forma de produto interno do vector $\mathbf{e}$ com as linhas das matrizes das derivadas de $\mathbf{Y}'\boldsymbol{\Theta}'^{-1}\left(\boldsymbol{I_T} \otimes \boldsymbol{\Sigma^{-1}}\right)$ com respeito ao parâmetro $\boldsymbol{\beta_i}$. Esses termos tem a forma de soma termos cruzados de t e ε_t's, multiplicados por combinações desfasadas de $Y_t's$ e ε_t's que convergem para zero em probabilidade quando divididos por T, devido ao facto de os termos desfasados de $Y_t's$ e ε_t's serem independentes do valor corrente de ε_t (Reinsel, 1997).

O método numérico iterativo pode ser implementado a partir da expressão (6.46)

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta_0} + \left[-\left(\frac{\partial^2 \boldsymbol{l}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\right)_{\boldsymbol{\beta_0}}\right]^{-1}\left(\frac{\partial \boldsymbol{l}}{\partial \boldsymbol{\beta}}\right)_{\boldsymbol{\beta_0}}$$

Em que $\boldsymbol{\beta_0}$ é a estimativa obtida na iteração anterior.

A estimativa inicial pode ser obtida, por exemplo, recorrendo à estimativa dos mínimos quadrados

$$\boldsymbol{\beta_0} = \left(\widetilde{\boldsymbol{\Phi}}_{\mathbf{1}}^{'}, \dots, \widetilde{\boldsymbol{\Phi}}_{\boldsymbol{p}}^{'}, \widetilde{\boldsymbol{\theta}}_{\mathbf{1}}^{'}, \dots, \widetilde{\boldsymbol{\theta}}_{\boldsymbol{q}}^{'}\right)^{'} = \boldsymbol{vec}\left(\widetilde{\boldsymbol{\Phi}}_{\mathbf{1}}, \dots, \widetilde{\boldsymbol{\Phi}}_{\boldsymbol{p}}, \widetilde{\Theta}_{\mathbf{1}}, \dots, \widetilde{\Theta}_{\boldsymbol{p}}\right)$$

Uma vez obtida a estimativa inicial, de (6.43) tem-se

$$\widetilde{\boldsymbol{\Theta}} = (\boldsymbol{I_T} \otimes \boldsymbol{I_k}) - \sum_{\boldsymbol{i=1}}^{\boldsymbol{q}} \left(\boldsymbol{L^i} \otimes \boldsymbol{\Theta_i}\right)$$

$$\tilde{\mathbf{e}} = \widetilde{\boldsymbol{\Theta}}^{-1}\left(\mathbf{z} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \left(\boldsymbol{\mathcal{B}^i}\mathbf{Z} \otimes \boldsymbol{I_k}\right)\widetilde{\boldsymbol{\Phi}}_i\right)$$

Recorrendo às equações (6.45) e (6.47), as equações modificadas de Newton-Raphson para $\widehat{\boldsymbol{\beta}}$ tem uma solução da forma

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta_0} + \underbrace{\left[\widetilde{\mathbf{Y}}'\widetilde{\boldsymbol{\Theta}}^{'-1}\left(\boldsymbol{I_T} \otimes \widetilde{\boldsymbol{\Sigma}}^{-1}\right)\widetilde{\boldsymbol{\Theta}}^{-1}\widetilde{\mathbf{Y}}\right]^{-1}}_{-\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\right)_{\boldsymbol{\beta_0}}} \underbrace{\widetilde{\mathbf{Y}}'\widetilde{\boldsymbol{\Theta}}^{'-1}\left(\boldsymbol{I_T} \otimes \widetilde{\boldsymbol{\Sigma}}^{-1}\right)\tilde{\mathbf{e}}}_{\left(\frac{\partial l}{\partial \boldsymbol{\beta}}\right)_{\boldsymbol{\beta_0}}}$$

Definindo-se $\overline{\mathbf{Y}} = \widetilde{\Theta}^{-1}\widetilde{\mathrm{Y}} = (\overline{\mathrm{Y}}_1, \overline{\mathrm{Y}}_2, \dots, \overline{\mathrm{Y}}_{\mathrm{T}})'$, com $\overline{\mathrm{Z}}_{\mathrm{t}} = (\partial\varepsilon_t^{'}/\partial\beta)_{\beta_0}$, e substituindo na expressão anterior tem-se finalmente (Reinsel, 1997)

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta_0} + \left[\overline{\mathbf{Y}}'\left(\boldsymbol{I_T} \otimes \widetilde{\boldsymbol{\Sigma}}^{-1}\right)\overline{\mathbf{Y}}\right]^{-1}\overline{\mathbf{Y}}'\left(\boldsymbol{I_T} \otimes \widetilde{\boldsymbol{\Sigma}}^{-1}\right)\tilde{\mathbf{e}} \tag{6.48}$$

$$= \boldsymbol{\beta_0} + \left[\sum_{\boldsymbol{t=1}}^{\mathbf{T}} \overline{\mathbf{Y}}_{\boldsymbol{t}}\widetilde{\boldsymbol{\Sigma}}^{-1}\overline{\mathbf{Y}}_{\boldsymbol{t}}^{'}\right]^{-1}\sum_{\boldsymbol{t=1}}^{\mathbf{T}} \overline{\mathbf{Y}}_{\boldsymbol{t}}\boldsymbol{\Sigma^{-1}}\boldsymbol{\varepsilon_t}$$

Com $\tilde{\varepsilon}_0 = \tilde{\varepsilon}_{-1} = \cdots = \tilde{\varepsilon}_{-q+1} = 0$.

Na prática, o procedimento iterativo (6-4) pode ter que ser modificado em certas alturas, por exemplo, pela introdução de um factor de ajustamento de escala do vector de incremento no membro direito de (6.48) para evitar a saída da zona de convergência e garantir o incremento da função, ou seja, que $|\Sigma|$ decresça em cada iteração. Ou seja, na j-ésima iteração, a estimativa deverá ser

$$\boldsymbol{\beta_j} = \boldsymbol{\beta_{j-1}} + \delta_j\left[\overline{\mathbf{Y}}_{j-1}^{'}\left(\boldsymbol{I_T} \otimes \widehat{\boldsymbol{\Sigma}}_{j-1}^{-1}\right)\overline{\mathbf{Y}}_{j-1}\right]^{-1}\overline{\mathbf{Y}}_{j-1}^{'}\left(\boldsymbol{I_T} \otimes \widehat{\boldsymbol{\Sigma}}_{j-1}^{-1}\right)\hat{\mathbf{e}}_{j-1}$$

onde δ_j é o factor de escala. Outro problema que pode ocorrer é que o operador MA estimado pode não ser invertível em determinadas iterações, ou seja $det\left[\widehat{\Theta}(\mathcal{B})\right] = 0$ pode ter algumas raízes menores ou iguais a um em valor absoluto, pelo que este operador poderá te que ser ajustado (Reinsel, 1997).

Quanto à distribuição assimptótica do estimador MLE de $\hat{\beta}$, de acordo com (Reinsel, 1997), verifica-se que

$$\boldsymbol{T^{1/2}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\xrightarrow{D}N\left(\mathbf{0},V^{-1}\right)} \qquad \text{quando } \boldsymbol{T}\to\infty \tag{6.49}$$

onde a notação $\xrightarrow{D}$ significa que "*converge em probabilidade*" e

$$\boldsymbol{T^{-1}\sum_{t=1}^{T}\overline{Y}_t\widetilde{\Sigma}^{-1}\overline{Y}_t'=T^{-1}\overline{Y}'\left(I_T\otimes\widetilde{\Sigma}^{-1}\right)\overline{Y}\xrightarrow{P}V} \qquad \text{quando } \boldsymbol{T}\to\infty \tag{6.50}$$

No caso de existirem restrições lineares, o estimador ML pode ser determinado através de iterações pelo método Newton-Raphson modificado similar a (6.48) dado por

$$\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma_0}+\left[\boldsymbol{R'\overline{Y}'}\left(\boldsymbol{I_T}\otimes\widetilde{\boldsymbol{\Sigma}}^{-1}\right)\boldsymbol{\overline{Y}R}\right]^{-1}\boldsymbol{R'\overline{Y}'}\left(\boldsymbol{I_T}\otimes\widetilde{\boldsymbol{\Sigma}}^{-1}\right)\tilde{\mathbf{e}} \tag{6.51}$$

onde $\overline{\mathbf{Y}}=\widetilde{\Theta}^{-1}\widetilde{\mathrm{Y}}$ e γ_0 é a estimativa de γ da iteração anterior.

Igualmente de forma similar, as propriedade assimptóticas são dadas por

$$\boldsymbol{V_\gamma = E\left[R'\overline{Y}'\Sigma^{-1}\overline{Y}R\right]=R'VR} \tag{6.52}$$

6.1.4.2 Determinação da ordem do modelo

Antes de se tentar determinar os parâmetros de um modelo ARMA(*p,q*), tem que se determinar a ordem dos respectivos operadores. Para o caso multivariado, um grande número de estratégias tem sido propostas, sem que nenhuma se tenha transformado em standard tal como aconteceu com a abordagem de Box-Jenkis para o caso univariado (Lütkepohl, 2007). Algumas estratégias são baseadas principalmente em subjectivas avaliações de certas características dos processos tais como a autocorrelação e autocorrelação parcial. A decisão das ordens específicas e restrições nos coeficientes das matrizes eram então baseadas nestas quantidades. Outros métodos utilizam uma mistura de testes estatísticos, o uso de critérios de selecção de modelo e a avaliação do analista. Outros ainda são baseados predominantemente em critérios estatísticos de selecção e, em princípio, podem ser determinados automaticamente por um computador. Os procedimentos automáticos têm a vantagem de as suas propriedades estatísticas podem ser determinadas rigorosamente. Nas aplicações actuais, são utilizadas frequentemente algumas de diferentes abordagens. Por outras palavras, a perícia e o conhecimento antecipado de um analista não deve ser preterido em favor dos procedimentos puramente estatísticos. Sugere-se que os modelos propostos por vários tipos de critérios e procedimentos sejam avaliados e desenvolvidos por peritos antes de se determinar o modelo final.

O modelo aqui proposto, e que foi utilizado ao longo deste trabalho é a generalização para modelos multivariados de uma técnica explorada por (Hannan & Rissanen, 1982) para modelos ARMA univariados.

No procedimento proposto, primeiro obtém-se uma estimativa da série dos resíduos (inovações) ε_t de um potencial modelo vectorial ARMA(*p,q*) pela aproximação do modelo a

um modelo autoregressivo puro de ordem suficientemente alta m^*. A ordem m^* do modelo autoregressivo aproximado pode ser escolhida recorrendo a critérios de selecção tais como o AIC, por exemplo, que sugere um valor de *m* para o qual o valor da função $\mathbf{log}\left(\left|\widetilde{\boldsymbol{\Sigma}}_{\mathbf{m}}\right|\right) + \boldsymbol{2mk^2/T}$ é minimizado. Com base no modelo seleccionado, AR(m^*) determinam-se os resíduos $\tilde{\boldsymbol{\varepsilon}}_{\boldsymbol{t}} = \boldsymbol{Y}_{\boldsymbol{t}} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{m^*}} \widehat{\boldsymbol{\Phi}}_{\mathbf{j}\boldsymbol{m^*}} \boldsymbol{Y}_{\boldsymbol{t-j}} - \widehat{\boldsymbol{\delta}}^*$, com $\boldsymbol{t} = \boldsymbol{m}^* + \mathbf{1}, \ldots, \boldsymbol{T}$. Num segundo estádio do procedimento faz-se a regressão linear de $\boldsymbol{Y}_{\boldsymbol{t}}$ em $\boldsymbol{Y}_{\boldsymbol{t-1}}, \ldots, \boldsymbol{Y}_{\boldsymbol{t-p}}$ e $\tilde{\boldsymbol{\varepsilon}}_{\boldsymbol{t-1}}, \ldots, \tilde{\boldsymbol{\varepsilon}}_{\boldsymbol{t-q}}$ para vários valores de p e q. Ou seja, estimam-se modelos da forma

$$\boldsymbol{Y}_{\boldsymbol{t}} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\Phi}_{\mathbf{j}} \boldsymbol{Y}_{\boldsymbol{t-j}} + \boldsymbol{\delta} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{q}} \boldsymbol{\Theta}_{\mathbf{j}} \tilde{\boldsymbol{\varepsilon}}_{\boldsymbol{t-j}} + \boldsymbol{\varepsilon}_{\boldsymbol{t}} \tag{6.53}$$

por regressão linear dos mínimos quadrados e determina-se a estimativa da matriz de covariância dos erros com base nos resíduos obtidos através da expressão anterior. Então, pela aplicação do critério BIC, a ordem (*p,q*) do modelo VARMA é determinada com base na no resultado que minimiza a função $\mathbf{log}\left(\left|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{p,q}}\right|\right) + (\boldsymbol{p}+\boldsymbol{q})\boldsymbol{k^2}(\boldsymbol{T})/\boldsymbol{T}$. O modelo obtido servirá de base para a primeira iteração do procedimento iterativo da estimativa de máxima verosimilhança descrito acima. O modelo pode subsequentemente ser validado recorrendo à análise de resíduos. A vantagem do procedimento é que a determinação da estimativa de máxima verosimilhança é bastante pesada em termos computacionais para se experimentar uma larga gama se valores de *p* e *q* por esse método. Adicionalmente, os parâmetros obtidos pelo método dos mínimos quadrados são geralmente excelentes valores de partida para as iterações de máxima verosimilhança.

6.1.4.3 Validação global da adequação do modelo

Normalidade dos resíduos

Um dos objectivos da modelação de processos poderá estar relacionado com o controlo estatístico multivariado, recorrendo ao controlo dos dados de saída do processo após filtragem pelo modelo então obtido, no pressuposto subjacente ao modelo, que se trata de dados normalmente distribuído (Figura 6-1).

A não normalidade dos resíduos indica geralmente que o modelo não é uma boa representação dos dados gerados pelo processo. Consequentemente, torna-se desejável que se teste a normalidade dos resíduos.

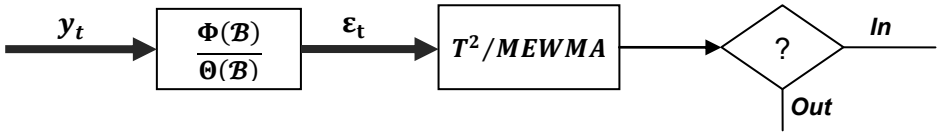


Figura 6-1 Aplicação de cartas de controlo com dados correlacionados multivariados

Os testes de normalidade multivariada utilizados neste trabalho são baseados nos terceiros e quartos momentos centrados da distribuição normal. Se x é uma variável aleatória univariada com distribuição normal estandardizada, $x \rightsquigarrow N(0,1)$, os seus terceiro e quarto

momento são dados por $E[x^3] = 0$ e $E[x^4] = 3$. Seja u_t um processo ruído branco de dimensão k com $u_t \rightsquigarrow N(\mu_u, \Sigma_u)$ e seja P uma matriz que satisfaz a relação $PP' = \Sigma_u$. P pode ser obtida por exemplo por uma decomposição de Choleski de Σ_u. Então

$$\boldsymbol{\omega_t = (\omega_{1t}, \dots, \omega_{kt})' = P^{-1}(u_t - \mu_u) \rightsquigarrow N(\mu_u, \Sigma_u)}$$

Ou seja, os componentes de ω_t são variáveis aleatórias independentes que seguem uma distribuição normal estandardizada. Consequentemente tem-se que

$$\boldsymbol{E\begin{bmatrix} \omega_{1t}^3 \\ \vdots \\ \omega_{Kt}^3 \end{bmatrix} = 0 \qquad \text{e} \qquad E\begin{bmatrix} \omega_{1t}^4 \\ \vdots \\ \omega_{Kt}^4 \end{bmatrix} = \begin{bmatrix} 3 \\ \vdots \\ 3 \end{bmatrix} = 3_K} \tag{6.54}$$

Este resultado será utilizado na validação da normalidade de processos "ruído branco".

Para construir o teste, assume-se de que se dispõe das observações $u_1, \dots, u_T$, a partir das quais se define

$$\boldsymbol{\bar{u} = \frac{1}{T}\sum_{t=1}^{T} u_t}$$

$$\boldsymbol{S_u = \frac{1}{T-1}\sum_{t=1}^{T}(u_t - \bar{u})(u_t - \bar{u})'}$$

e P_s é uma matriz para a qual $P_s P_s' = S_u$ e tal que o limite de probabilidade de $(P_s - P)$ é zero. Define-se ainda

$$\boldsymbol{v_t = (v_{1t}, \dots, v_{kt})' = P_s^{-1}(u_t - \bar{u}), \qquad t = 1, \dots, T}$$

$$\boldsymbol{b_1 = (b_{11}, \dots, b_{k1})' \quad \text{com} \quad b_{k1} = \frac{1}{T}\sum_t v_{kt}^3 \quad k = 1, \dots, K} \tag{6.55}$$

$$\boldsymbol{b_2 = (b_{12}, \dots, b_{k2})' \quad \text{com} \quad b_{k2} = \frac{1}{T}\sum_t v_{kt}^4 \quad k = 1, \dots, K} \tag{6.56}$$

Deste modo, b_1 e b_2 são estimadores dos vectores (6.54). Após estas definições, se u_t é ruído branco Gaussiano com matriz de covariância não singular $\boldsymbol{\Sigma_u}$ e valor esperado μ_u, $\boldsymbol{u_t \rightsquigarrow N(\mu_u, \Sigma_u)}$, então verifica-se que (Lütkepohl, 2007)

$$\boldsymbol{\sqrt{T}\begin{bmatrix} b_1 \\ b_2 - 3_K \end{bmatrix} \xrightarrow{d} N\left(0, \begin{bmatrix} 6I_K & 0 \\ 0 & 24I_K \end{bmatrix}\right)} \tag{6.57}$$

ou seja, b_1 e b_2 são assimptoticamente independentes e normalmente distribuídos, (6.57) implica que

$$\boldsymbol{\lambda_s = T b_1' b_1 / 6 \xrightarrow{d} \chi^2(K)} \tag{6.58}$$

e

$$\boldsymbol{\lambda_k = T(b_2 - 3_K)'(b_2 - 3_K)/24 \xrightarrow{d} \chi^2(K)} \tag{6.59}$$

Para a primeira estatística pode definir-se o teste de hipótese

$$\boldsymbol{H_0 : E\begin{bmatrix} \omega_{1t}^3 \\ \vdots \\ \omega_{Kt}^3 \end{bmatrix} = 0} \qquad \boldsymbol{H_1 : E\begin{bmatrix} \omega_{1t}^3 \\ \vdots \\ \omega_{Kt}^3 \end{bmatrix} \neq 0} \tag{6.60}$$

Para a segunda estatística pode definir-se o teste de hipótese

$$\boldsymbol{H_0 : E\begin{bmatrix} \omega_{1t}^4 \\ \vdots \\ \omega_{Kt}^4 \end{bmatrix} = 3_K} \qquad \boldsymbol{H_1 : E\begin{bmatrix} \omega_{1t}^4 \\ \vdots \\ \omega_{Kt}^4 \end{bmatrix} \neq 3_K} \tag{6.61}$$

Adicionalmente tem-se

$$\boldsymbol{\lambda_{sk} = \lambda_s + \lambda_k \xrightarrow{d} \chi^2(2K)} \tag{6.62}$$

Que pode ser utilizada para um teste conjunto de (6.60) e (6.61).

Matrizes de correlação dos resíduos

Um teste global para determinar a adequação do modelo obtido, por MLE ou LSE, pode ser feito com recurso à análise das matrizes de autocorrelação dos resíduos. Uma vez estimados os parâmetros de um modelo ARMA(p,q), os resíduos podem ser determinados através da expressão

$$\boldsymbol{\hat{\varepsilon}_t = Z_t - \sum_{j=1}^{p} \hat{\Phi}_j Z_{t-j} - \hat{\delta} + \sum_{j=1}^{q} \Theta_j \hat{\varepsilon}_{t-j} \qquad t = 1, \ldots, T}$$

Os elementos individuais das matrizes de correlação e possíveis padrões deverão ser avaliados no sentido em que os resíduos devem ter um comportamento de acordo com um processo ruído branco. Os limites de dois desvios padrão, $\pm 2/\sqrt{T}$ poderão ser utilizados para garantir a significância dos das correlações residuais individuais.

Adicionalmente, pode-se aplicar um teste global de ajustamento a uma sequência de matrizes de correlação residual, teste de "portmanteau" dado por (Isermann, 1989)

$$\begin{aligned} \boldsymbol{Q_s} &= \boldsymbol{T^2 \sum_{l=1}^{s} (T-l)^{-1} \sum_{i=1}^{k} \sum_{j=1}^{k} r_{ij}(l) \sum_{j=1}^{k} r_{ji}(-l)} \\ &= \boldsymbol{T^2 \sum_{l=1}^{s} (T-l)^{-1} tr\{C_\varepsilon(l)\hat{\Sigma}^{-1} C_\varepsilon(-l)\hat{\Sigma}^{-1}\}} \end{aligned} \tag{6.63}$$

ou equivalentemente por

$$\boldsymbol{Q_s = T^2 \sum_{l=1}^{s} (T-l)^{-1}\, tr\{\hat{\rho}_\varepsilon(l)\hat{\rho}_\varepsilon(0)^{-1}\hat{\rho}_\varepsilon(-l)\hat{\rho}_\varepsilon(0)^{-1}\}} \quad (6.64)$$

De acordo com a hipótese nula em que o processo segue um processo VARMA(p,q), a estatística de teste tem distribuição aproximada χ^2 com $k^2(s-p-q)$ graus de liberdade. Consequentemente, o modelo é rejeitado para grandes valores de Q_s. A distribuição aproximada é valida no pressuposto de que s é suficientemente grande de modo a que as matrizes Ψ_j da representação MA infinita para o modelo ARMA são insignificantes para $j > s$.

6.1.5 Modelos de cointegração e de rank reduzido

6.1.5.1 Modelos de rank reduzido e análise de correlação canónica parcial

Nesta secção considera-se modelos autoregressivos multivariados AR(p), que poderão ser representados por uma estrutura de rank reduzido nos seus coeficientes matriciais Φ_j. Mais especificamente, considera-se modelos AR(p) do tipo

$$\boldsymbol{Z_t = \sum_{j=1}^{p} \Phi_j Z_{t-j} + \delta + \varepsilon_t} \quad (6.65)$$

Neste modelo assume-se que as matrizes Φ_j têm uma estrutura particular de rank reduzido tal que $rank(\Phi_j) = r_j \geq rank(\Phi_{j+1}) = r_{j+1}$ com $j = 1, 2, \ldots, p-1$. Deste modo, os Φ_j podem ser representados na forma $\Phi_j = A_j B_j$, onde A_j e B_j são matrizes de dimensão $k \times r_j$ e $r_j \times k$ respectivamente. Assim a equação anterior pode ser escrita na forma

$$\boldsymbol{Z_t = \sum_{j=1}^{p} A_j B_j Z_{t-j} + \delta + \varepsilon_t}$$

Estes modelos podem resultar em parametrizações mais parcimoniosas, estruturas mais detalhadas e possivelmente simplificadas, e possivelmente em interpretações mais úteis e interessantes no que diz respeito às inter-relações entre as k séries temporais.

Para se obter a informação sobre o rank das matrizes Φ_j nestes modelos, recorre-se normalmente à análise das correlações canónicas parciais. Uma das consequências fundamentais destes modelos é que existirão no mínimo $k - r_j$ correlações canónicas parciais nulas entre Z_t e Z_{t-j}, dado $Z_{t-1}, \ldots, Z_{t-j+1}$. Ou seja, pode-se determinar uma matriz F'_j de dimensão $(k - r_j) \times k$ cujas linhas são linearmente independentes tal que $F'_j A_j = \mathbf{0}$ e, consequentemente, $F'_j A_i = \mathbf{0}$ para $i > j$. Assim, recorrendo-se à análise das correlações canónicas parciais para vários valores de $j = 1, 2, \ldots$ pode-se identificar a estrutura do rank para o modelo, bem como a ordem global *p* do modelo AR(*p*).

A estatística de teste que pode ser utilizada para (por tentativa) especificar o rank é dada por

$$\boldsymbol{C(j,s) = -(T-j-1)\sum_{i=(k-s)+1}^{k} log\left(1-\hat{\rho}_i^2(j)\right)} \quad \textbf{(6.66)}$$

Para $s = 1, 2, \dots, k$, onde $1 \geq \hat{\rho}_1(j) \geq \hat{\rho}_2(j) \geq \cdots \geq \hat{\rho}_i(j) \geq 0$ são as estimativas amostrais das correlações canónicas parciais entre $\boldsymbol{Z_t}$ e $\boldsymbol{Z_{t-j}}$, dado $\boldsymbol{Z_{t-1}, \dots, Z_{t-j+1}}$ (ver secção 6.1.2.4). Nas condições de hipótese nula de que $rank(\boldsymbol{\Phi_j}) \leq k - s$ dentro da estrutura de modelo de rank reduzido encadeado, a estatística $\boldsymbol{C(j,s)}$ tem distribuição é assimptótica χ^2 com s^2 graus de liberdade (Reinsel, 1997). Assim, se o valor da estatística de teste não é "demasiado grande" não se deve rejeitar a hipótese nula.

6.1.5.2 Modelos autoregressivos multivariados não estacionários

Como se viu anteriormente, na modelação de séries temporais univariadas, é prática comum diferenciar as séries quando estas apresentam algumas características de não estacionaridade. A decisão no que diz respeito à necessidade de diferenciação é por vezes baseada, de forma informal, nas características da representação gráfica das séries e nas suas funções de autocorrelação parcial amostral. Esta situação tem conduzido a um interesse em estabelecer procedimentos de inferência mais formal na determinação da ordem de diferenciação apropriada das séries.

Para um simples processo univariado de primeira ordem AR(1) dado por $Y_t = \Phi Y_{t-1} + \varepsilon_t$, tem-se que o estimador dos mínimos quadrados de Φ é dado por

$$\boldsymbol{\hat{\Phi} = \sum_{t=1}^{T} Y_{t-1}Y_t \Big/ \sum_{t=1}^{T} Y_{t-1}^2 = \Phi + \sum_{t=1}^{T} Y_{t-1}\varepsilon_t \Big/ \sum_{t=1}^{T} Y_{t-1}^2}$$

Perante um processo AR(*p*), descrito por $\Phi(\mathcal{B})Y_t = \varepsilon_t$, ao colocar-se a hipótese de que o processo é não estacionário e que terá de ser diferenciado a fim de se obter um processo estacionário, corresponde a colocar-se a hipótese de que $\Phi(\mathcal{B}) = \Phi^*(\mathcal{B})(1-\mathcal{B})$, em que $\Phi^*(\mathcal{B})$ é um operador autoregressivo estacionário de ordem $p-1$. Nesse caso ter-se-á

$$\boldsymbol{\Phi(\mathcal{B})Y_t = \Phi^*(\mathcal{B})(1-\mathcal{B})Y_t = Y_t - Y_{t-1} - \sum_{j=1}^{p-1} \Phi_j^*\left(Y_{t-j} - Y_{t-j-1}\right)}$$

Fazer o teste de que $\Phi(\mathcal{B})$ tem uma raiz unitária é equivalente a colocar a hipótese de que $\rho = 1$ para o modelo $Y_t = \rho Y_{t-1} - \sum_{j=1}^{p-1} \Phi_j^*\left(Y_{t-j} - Y_{t-j-1}\right) + \varepsilon_t$, ou fazer o teste de que $\rho - 1 = 0$ para o modelo

$$\boldsymbol{Y_t - Y_{t-1} = (\rho - 1)Y_{t-1} - \sum_{j=1}^{p-1} \Phi_j^*\left(Y_{t-j} - Y_{t-j-1}\right) + \varepsilon_t}$$

De facto, definindo $W_t = Y_t - Y_{t-1}$, é fácil de verificar que qualquer modelo AR(p) $Y_t = \sum_{j=1}^{p} \Phi_j Y_{t-j} + \varepsilon_t$ pode ser expresso na forma equivalente

$$\boldsymbol{W_t = (\rho - 1)Y_{t-1} + \sum_{j=1}^{p-1} \Phi_j^* \left(Y_{t-j} - Y_{t-j-1}\right) + \varepsilon_t}$$

onde $\rho - 1 = \Phi(1) = \sum_{j=1}^{p} \Phi_j - 1$ e $\Phi_j^* = -\sum_{i=j+1}^{p} \Phi_i$, pelo que, dizer que existe uma raiz unitária no operador AR(*p*) é equivalente a afirmar que $\rho = \sum_{j=1}^{p} \Phi_j = 1$. Deste modo, seja $\left(\hat{\rho},\ \hat{\Phi}_1^*,\ \ldots\ , \hat{\Phi}_{p-1}^*\right)$ os coeficientes estimados do modelo por regressão linear dos mínimos quadrados obtidos pela regressão de Y_t em Y_{t-1}, W_{t-1}, … , W_{t-p+1}. Então, sob a hipótese de um modelo de raiz unitária onde $\rho = 1$ e de que o processo $\Phi^*(\mathcal{B}) = 1 - \sum_{j=1}^{p-1} \Phi_j^* \mathcal{B}^j$ é estacionário, tem-se que $(\hat{\rho} - 1) \Big/ \left\{ S_\varepsilon \left(\sum_{t=p+1}^{T} Y_{t-1}^2 \right)^{-1/2} \right\}$ tem distribuição limite da estatística $\hat{\tau}$ dada pela expressão

$$\boldsymbol{\hat{\tau} \xrightarrow{D} \int_0^1 B(u) dB(u) \Big/ \left(\int_0^1 B(u)^2 du \right)^{1/2}}$$

onde $B(\boldsymbol{u})$ é movimento Browniano padrão *ou processo de Wiener* no intervalo $[0, 1]$.

6.2 Modelos lineares com variáveis exógenas

Este será talvez a secção fulcral deste trabalho. Ao longo das secções anteriores do ponto 6, assumiu-se que todas variáveis do sistema dinâmico são determinadas conjuntamente dentro do sistema e de que todas tinham o mesmo status no modelo multivariado. Na prática, as variáveis de saída de um sistema dinâmico, denominada por exemplo por $Y_t = (Y_{1t}, \ldots\ , Y_{kt})'$, podem ser influenciadas por outras variáveis, denominadas como variáveis de entrada $X_t = (X_{1t}, \ldots\ , X_{kt})'$, que podem ser determinadas independentemente do sistema. Essas variáveis, denominadas como variáveis exógenas, podem ter tipos e origens diversas, sendo umas consideradas manipuláveis, uma vez que as podemos ajustar de modo a controlar as variáveis de saída, e outras restritivas, uma vez que não as podemos manipular mas influenciam igualmente as variáveis de saída.

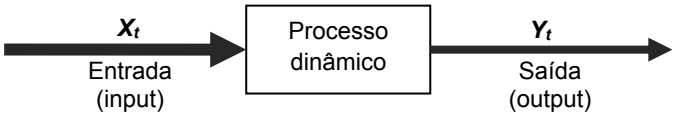


Figura 6-2 Processo dinâmico Multivariado

Em contraste, às variáveis determinadas dentro do sistema dá-se o nome de variáveis endógenas.

6.2.1 Tipos de Representação

Existem vários tipos de modelos para representar a influência das variáveis exógenas de entrada $\boldsymbol{X_t}$ nas variáveis de saída $\boldsymbol{Y_t}$, tais como função de transferência múltipla, modelos de equações dinâmicas simultâneas, modelos ARMAX vectoriais. O modelo de *função de transferência múltipla* é dado pela forma

$$\boldsymbol{Y_t} = \sum_{\boldsymbol{j=0}}^{\infty} \boldsymbol{\Psi_j^*}\, \boldsymbol{X_{t-j}} + \boldsymbol{N_t} \tag{6.67}$$

Onde Ψ_j^* são as matrizes coeficiente de dimensão $k \times r$ correspondentes à resposta ao impulso e N_t é um vector de ruído de dimensão k. Para se ter uma ideia mais intuitiva da natureza das matrizes Ψ_j^*, suponha-se que o processo se encontra no estado estacionário com todas as entradas e saídas inicialmente a zero. Aplicando-se um pulso unitário no instante zero na variável de entrada j:

$$\boldsymbol{X_j(0) = 1} \qquad \boldsymbol{X_j(t) = 0} \qquad \textbf{para } \boldsymbol{t} > 0$$

Seja a sequência de respostas da saída i dadas por

$$\left(\boldsymbol{\psi_{ij}(0)}\,,\, \boldsymbol{\psi_{ij}(1)},\, \boldsymbol{\psi_{ij}(2)},\ \ldots\right)$$

Tal que o vector das respostas de todas as saídas no instante t é

$$\left[\boldsymbol{\psi_{1j}(t)}\,,\, \boldsymbol{\psi_{2j}(t)},\ \ldots\,,\, \boldsymbol{\psi_{kj}(t)}\right]^{\boldsymbol{T}}$$

Pode-se então construir uma matriz que, no instante t, mostre como cada variável de saída responde a cada impulso unitário de cada variável de entrada

$$\boldsymbol{\Psi^*(t)} = \begin{bmatrix} \boldsymbol{\psi_{11}(t)} & \boldsymbol{\psi_{12}(t)} & \cdots & \boldsymbol{\psi_{1r}(t)} \\ \boldsymbol{\psi_{21}(t)} & \boldsymbol{\psi_{22}(t)} & \cdots & \boldsymbol{\psi_{2r}(t)} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\psi_{k1}(t)} & \boldsymbol{\psi_{k2}(t)} & \cdots & \boldsymbol{\psi_{kr}(t)} \end{bmatrix}$$

A matriz N_t é um vector de ruído de dimensão k que se assume poder ser representado por um processo ARMA, $\boldsymbol{\Phi(\mathcal{B})N_t = \Theta(\mathcal{B})\varepsilon_t}$. A variável exógena $\{X_t\}$ pode ser constituída por variáveis estocásticas e/ou determinísticas ou mesmo incluir variáveis com padrão sazonal. De notar que a expressão (6.67) pode ser vista como generalização para a análise multivariada da função de transferência Box-Jenkis.

Outro modelo, com maior relevância neste trabalho, com variáveis de entrada exógenas é o denominado modelo vectorial ARMAX, ou modelo VARMAX. Esta forma pode ser determinada a partir d a expressão (6.67) assumindo que o operador da função de transferência $\boldsymbol{\Psi^*(\mathcal{B})} = \sum_{j=0}^{\infty} \boldsymbol{\Psi_j^*\, \mathcal{B}^j}$ pode ser representado na forma racional $\boldsymbol{\Psi^*(\mathcal{B})} = \boldsymbol{\Phi(\mathcal{B})^{-1}\Theta^*(\mathcal{B})} = \sum_{j=0}^{\infty} \boldsymbol{\Psi_j^*\, \mathcal{B}^j}$ onde $\boldsymbol{\Theta^*(\mathcal{B})} = \sum_{j=0}^{s} \boldsymbol{\Theta_j^*\, \mathcal{B}^j}$ é de ordem s e $\sum_{j=0}^{s} \boldsymbol{\Theta_j^*\, \mathcal{B}^j}$ são matrizes de dimensão $k \times r$. Por conveniência assume-se que o factor $\boldsymbol{\Phi(\mathcal{B})}$ em $\boldsymbol{\Psi^*(\mathcal{B})}$ é o mesmo que o factor autoregressivo $\boldsymbol{\Phi(\mathcal{B})}$ no modelo para as perturbações $\boldsymbol{N_t}$, ou seja, considera-se dinâmicas idênticas para o sistema e para as perturbações. Deste modo, o modelo ARMAX pode ser expresso na forma

$$\boldsymbol{Y_t} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\Phi_j\, Y_{t-j}} = \sum_{\boldsymbol{j=0}}^{\boldsymbol{s}} \boldsymbol{\Theta_j^*\, X_{t-j}} + \boldsymbol{\varepsilon_t} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{q}} \boldsymbol{\Theta_j\, \varepsilon_{t-j}} \tag{6.68}$$

Esta expressão também pode ser escrita noutra forma em que envolve uma matriz coeficiente (triangular inferior) para Y_t ou ε_t na lag zero, reflectindo assim ligações instantâneas entre as variáveis endógena. Estas formas são normalmente referidas como modelo de equações dinâmicas simultâneas (*dynamic simultaneous equations model*).

De forma similar aos modelos vectoriais (ou multivariados) ARMA, o modelo ARMAX **(6.68**) diz-se estável se todas as raízes de $det\{\boldsymbol{\Phi}(\boldsymbol{\mathcal{B}})\} = 0$ são todas maiores que um em valor absoluto. Nessas condições, se a série de entrada $\{\boldsymbol{X_t}\}$ é um processo estacionário, bem como $\{\boldsymbol{\varepsilon_t}\}$**,** então a saída $\{\boldsymbol{Y_t}\}$ será estacionária. O processo terá então uma representação convergente dada por

$$\boldsymbol{Y_t} = \boldsymbol{\Psi^*(\mathcal{B})X_t} + \boldsymbol{\Psi(\mathcal{B})\varepsilon_t} = \sum_{\boldsymbol{j=0}}^{\infty} \boldsymbol{\Psi_j^*\, X_{t-j}} + \sum_{\boldsymbol{j=0}}^{\infty} \boldsymbol{\Psi_j\, \varepsilon_{t-j}} \tag{6.69}$$

onde $\boldsymbol{\Psi(\mathcal{B})} = \boldsymbol{\Phi(\mathcal{B})^{-1}\Theta(\mathcal{B})} = \sum_{\boldsymbol{j=0}}^{\infty} \boldsymbol{\Psi_j\, \mathcal{B}^j}$. Como já foi referido, as matrizes coeficiente Ψ_j^* representam os efeitos que as alterações nas variáveis de entrada provocam nas variáveis de saída nas várias *lags* de tempo e são conhecidas como matrizes resposta ao impulso. O ganho total do sistema dinâmico, ou ganho estacionário, é dado pelos elementos da matriz resposta ao degrau unitário $G = \Psi^*(1) = \sum_{j=0}^{\infty} \Psi_j^*$.

6.2.2 Previsão com modelos vectoriais ARMAX

Como já foi referido anteriormente, as variáveis exógenas não tem todas as mesmas características. A característica comum é que todas elas entram no processo, e influenciam as variáveis de saída ou resposta. Podem existir variáveis de entrada que não são manipuláveis, apesar de ser conhecido o seu níveis nos diversos instantes de tempo. Poderão existir outras que, apesar de serem manipuláveis, tem alguma dinâmica, pelo que o resultado das manipulações efectuadas no instante t, ou o efeito que as alterações noutras variáveis exógenas no instante t têm nessa variável, só se verificará alguns instantes mais tarde. Existem ainda as variáveis exógenas manipuláveis, no sentido em que o nível da variável pode ser alterado instantaneamente, ou que a sua constante de tempo é insignificante em relação à constante de tempo do sistema.

6.2.2.1 Previsão com variáveis exógenas estocásticas com dinâmica

Quando a série exógena $\{\boldsymbol{X_t}\}$, no modelo ARMAX (6.68) é estocástica e os valores futuros são desconhecidos, então esses valores terão que ser também eles previstos para prever os valores das respostas futuras $\boldsymbol{Y_t}$. Nesta situação assume-se que $\boldsymbol{X_t}$ é gerado por um processo multivariado ARMA de ordem *p* e *q*

$$\boldsymbol{X_t} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{A_j\, X_{t-j}} = \boldsymbol{a_t} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{q}} \boldsymbol{C_j\, a_{t-j}} \tag{6.70}$$

onde, por conveniência, se assume que $\{a_t\}$ e $\{\varepsilon_t\}$ são processos ruído brando independentes. Nestas condições, a melhor previsão (matriz MSE mínima) l períodos à frente de Y_{t+l}, com base na informação disponível no instante t será

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{t}}(\boldsymbol{l}) = \boldsymbol{E}[\boldsymbol{Y}_{\boldsymbol{t+l}}|\boldsymbol{Y}_{\boldsymbol{t}}, \boldsymbol{Y}_{\boldsymbol{t-1}}, \dots, \boldsymbol{X}_{\boldsymbol{t}}, \boldsymbol{X}_{\boldsymbol{t-1}}, \dots]$$

A previsão óptima l períodos à frente pode ser expressa através da expressão recursiva (Reinsel, 1997)

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{t}}(\boldsymbol{l}) = \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\Phi}_{\boldsymbol{j}}\, \widehat{\boldsymbol{Y}}_{\boldsymbol{t}}(\boldsymbol{l-j}) + \sum_{\boldsymbol{j=0}}^{\boldsymbol{s}} \boldsymbol{\Theta}_{\boldsymbol{j}}^{*}\, \widehat{\boldsymbol{X}}_{\boldsymbol{t}}(\boldsymbol{l-j}) - \sum_{\boldsymbol{j=l}}^{\boldsymbol{q}} \boldsymbol{\Theta}_{\boldsymbol{j}}\, \boldsymbol{\varepsilon}_{\boldsymbol{t+l-j}} \tag{6.71}$$

Para $l = 1, \dots, q$ e

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{t}}(\boldsymbol{l}) = \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\Phi}_{\boldsymbol{j}}\, \widehat{\boldsymbol{Y}}_{\boldsymbol{t}}(\boldsymbol{l-j}) + \sum_{\boldsymbol{j=0}}^{\boldsymbol{s}} \boldsymbol{\Theta}_{\boldsymbol{j}}^{*}\, \widehat{\boldsymbol{X}}_{\boldsymbol{t}}(\boldsymbol{l-j}) \quad \textbf{para} \;\; \boldsymbol{l} > q$$

Em que obviamente $\hat{Y}_t(l-j) \equiv \boldsymbol{Y_{t+l-j}}$ e $\widehat{\boldsymbol{X}}_t(l-j) \equiv \boldsymbol{X_{t+l-j}}$ para $l \leq j$, e em que, também obviamente, $\hat{X}_t(l-j)$ é dado pela expressão

$$\widehat{\boldsymbol{X}}_{\boldsymbol{t}}(\boldsymbol{l}) = \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \mathbf{A}_{\boldsymbol{j}}\, \widehat{\boldsymbol{X}}_{\boldsymbol{t}}(\boldsymbol{l-j}) - \sum_{\boldsymbol{j=l}}^{\boldsymbol{q}} \mathbf{C}_{\boldsymbol{j}}\, \boldsymbol{a}_{\boldsymbol{t+l-j}}$$

Uma propriedade que é conveniente verificar na fase de modelação é

$$\widehat{\boldsymbol{X}}_{\boldsymbol{t}}(\boldsymbol{l}) = \boldsymbol{E}[\boldsymbol{X}_{\boldsymbol{t+l}}|, \boldsymbol{X}_{\boldsymbol{t}}, \boldsymbol{X}_{\boldsymbol{t-1}}, \dots, \boldsymbol{Y}_{\boldsymbol{t}}, \boldsymbol{Y}_{\boldsymbol{t-1}}, \dots] \equiv \boldsymbol{E}[\boldsymbol{X}_{\boldsymbol{t+l}}|, \boldsymbol{X}_{\boldsymbol{t}}, \boldsymbol{X}_{\boldsymbol{t-1}}, \dots] \tag{6.72}$$

Quando esta propriedade se verifica diz-se que Y_t não causa $\boldsymbol{X_t}$ ou que não existe feedback de Y_t para $\boldsymbol{X_t}$.

Outra forma de obter previsões para o modelo ARMAX é combinar as variáveis exógenas com as variáveis endógenas numa única variável vectorial $\boldsymbol{Z_t} = (\boldsymbol{X_t'}, \boldsymbol{Y_t'})$e a partir dela definir um processo ARMA. De (6.68) e (6.70) verifica-se que $\boldsymbol{Z_t}$ satisfaz o modelo ARMA

$$\begin{bmatrix} \boldsymbol{I_r} & \boldsymbol{0} \\ -\boldsymbol{\Theta_0^*} & \boldsymbol{I_r} \end{bmatrix} \begin{bmatrix} \boldsymbol{X_t} \\ \boldsymbol{Y_t} \end{bmatrix} - \begin{bmatrix} \boldsymbol{A_1} & \boldsymbol{0} \\ \boldsymbol{\Theta_1^*} & \boldsymbol{\Phi_1} \end{bmatrix} \begin{bmatrix} \boldsymbol{X_{t-1}} \\ \boldsymbol{Y_{t-1}} \end{bmatrix} - \dots - \begin{bmatrix} \boldsymbol{A_p} & \boldsymbol{0} \\ \boldsymbol{\Theta_p^*} & \boldsymbol{\Phi_p} \end{bmatrix} \begin{bmatrix} \boldsymbol{X_{t-p}} \\ \boldsymbol{Y_{t-p}} \end{bmatrix}$$
$$= \begin{bmatrix} \boldsymbol{a_t} \\ \boldsymbol{\varepsilon_t} \end{bmatrix} - \begin{bmatrix} \boldsymbol{C_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Theta_1} \end{bmatrix} \begin{bmatrix} \boldsymbol{a_{t-1}} \\ \boldsymbol{\varepsilon_{t-1}} \end{bmatrix} - \dots - \begin{bmatrix} \boldsymbol{C_q} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Theta_q} \end{bmatrix} \begin{bmatrix} \boldsymbol{a_{t-q}} \\ \boldsymbol{\varepsilon_{t-q}} \end{bmatrix} \tag{6.73}$$

Multiplicando tudo pela matriz inversa da *lag* zero, obtém-se um processo ARMA multivariado na forma standard. A característica básica deste modelo é que ambos os operadores MA e AR são compostos por matrizes bloco - triangulares inferiores. Este facto reflecte a característica exógena das variáveis X_t e a não causalidade de $\boldsymbol{Y_t}$ para X_t. Se $\Theta_0^* = 0$ em (6.68) e (6.70), e consequentemente em (6.73), as variáveis $\boldsymbol{Y_t}$ e X_t dizem-se *instantaneamente não causal*, e então tem-se

$$\boldsymbol{E}[\boldsymbol{Y}_{\boldsymbol{t+1}}|, \boldsymbol{X}_{\boldsymbol{t+1}}, \boldsymbol{X}_{\boldsymbol{t}}, \boldsymbol{X}_{\boldsymbol{t-1}}, \dots, \boldsymbol{Y}_{\boldsymbol{t}}, \boldsymbol{Y}_{\boldsymbol{t-1}}, \dots]$$
$$\equiv \boldsymbol{E}[\boldsymbol{Y}_{\boldsymbol{t+1}}|, \boldsymbol{X}_{\boldsymbol{t}}, \boldsymbol{X}_{\boldsymbol{t-1}}, \dots, \boldsymbol{Y}_{\boldsymbol{t}}, \boldsymbol{Y}_{\boldsymbol{t-1}}, \dots]$$

6.2.2.2 Matriz do erro quadrático médio (MSE) de previsão óptima

Os erros de previsão cometidos pelo recurso à expressão (6.71) poderão ser determinados através da expressão

$$\boldsymbol{e_t(l) = Y_{t+l} - \hat{Y}_t(l)} \tag{6.74}$$

De (6.69) e (6.70) tem-se que o processo $\boldsymbol{Y_t}$ também pode adquirir a forma (MA infinita)

$$\boldsymbol{Y_t = \sum_{j=0}^{\infty} V_j\, a_{t-j} + \sum_{j=0}^{\infty} \Psi_j\, \varepsilon_{t-j}} \tag{6.75}$$

Em que se definiu

$$\boldsymbol{V(\mathcal{B}) = \sum_{j=0}^{\infty} V_j\, \mathcal{B}^j = \Psi^*(\mathcal{B})\Psi_x(\mathcal{B})}$$

com $\Psi^*(\mathcal{B}) = \boldsymbol{\Phi(\mathcal{B})^{-1}\Theta^*(\mathcal{B})}$ e $\Psi_x(\mathcal{B}) = \boldsymbol{A(\mathcal{B})^{-1}C(\mathcal{B})}$. Consequentemente tem-se que $\hat{Y}_t(l)$ pode ser determinado a partir da expressão

$$\boldsymbol{\hat{Y}_t(l) = \sum_{j=l}^{\infty} V_j\, a_{t+l-j} + \sum_{j=l}^{\infty} \Psi_j\, \varepsilon_{t+l-j}}$$

Pelo que

$$\begin{aligned}
\boldsymbol{e_t(l)} &= \boldsymbol{Y_{t+l} - \hat{Y}_t(l)} \\
\boldsymbol{e_t(l)} &= \left[\sum_{j=0}^{\infty} \mathbf{V}_j\, \boldsymbol{a_{t+l-j}} + \sum_{j=0}^{\infty} \boldsymbol{\Psi_j\, \varepsilon_{t+l-j}}\right] - \left[\sum_{j=l}^{\infty} \mathbf{V}_j\, \boldsymbol{a_{t+l-j}} + \sum_{j=l}^{\infty} \boldsymbol{\Psi_j\, \varepsilon_{t+l-j}}\right] \\
&= \sum_{j=0}^{l-1} \mathbf{V}_j\, \boldsymbol{a_{t+l-j}} + \sum_{j=0}^{l-1} \boldsymbol{\Psi_j\, \varepsilon_{t+l-j}}
\end{aligned}$$

Relembrando que $E[a_t a_t^{'}] = Cov(a_t) = \Sigma_a$, $E[\varepsilon_t \varepsilon_t^{'}] = Cov(\varepsilon_t) = \Sigma_\varepsilon$ e que $E[a_t \varepsilon_t^{'}] = 0$ tem-se que a matriz MSE dos erros de previsão l períodos à frente será dada por

$$\boldsymbol{\Sigma(l) = E[e_t(l) e_t^{'}(l)] = \sum_{j=0}^{l-1} V_j\, \Sigma_a V_j^{'} + \sum_{j=0}^{l-1} \Psi_j\, \Sigma_\varepsilon \Psi_j^{'}} \tag{6.76}$$

6.2.2.3 Previsão com variáveis exógenas especificadas

A previsão para futuros valores de $\boldsymbol{Y_t}$ na secção anterior é baseada na situação de que os valores futuros da variável exógena são estocásticos e tem de ser também eles previstos.

Noutras ocasiões, ou mesmo dentro do mesmo processo, futuros valores de $\boldsymbol{X_t}$ podem ser previamente conhecidos, por exemplo, variáveis manipuláveis em que o nível da variável pode-se altera de forma determinística e instantânea, ou outra variáveis, que apesar de não serem manipuláveis, ou seu valor é de alguma forma conhecido. Nesses casos pode-se estar interessados em prever futuros valores de $\boldsymbol{Y_t}$ condicional aos valores futuros da variável exógena $\boldsymbol{X_t}$ e na matriz MSE dos erros de previsão.

Do modelo ARMAX (6.68) conclui-se facilmente que a previsão óptima $\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}(\boldsymbol{l})$ de $\boldsymbol{Y_{t+l}}$, condicionada às informações passadas e também ao conhecimento futuro dos valores $\boldsymbol{X_{t+1}}, \ldots, \boldsymbol{X_{t+l}}$ satisfaz a relação

$$\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}(\boldsymbol{l}) = \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\Phi_j}\, \widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}(\boldsymbol{l-j}) + \sum_{\boldsymbol{j=0}}^{\boldsymbol{s}} \boldsymbol{\Theta_j^*}\, \boldsymbol{X_{t+l-j}} - \sum_{\boldsymbol{j=l}}^{\boldsymbol{q}} \boldsymbol{\Theta_j}\, \boldsymbol{\varepsilon_{t+l-j}} \tag{6.77}$$

Para $l = 1, \ldots, q$ e

$$\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}(\boldsymbol{l}) = \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\Phi_j}\, \widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}(\boldsymbol{l-j}) + \sum_{\boldsymbol{j=0}}^{\boldsymbol{s}} \boldsymbol{\Theta_j^*}\, \boldsymbol{X_{t+l-j}} \quad \textbf{para} \;\; \boldsymbol{l} > q$$

Em que obviamente $\tilde{Y}_t(l-j) \equiv \boldsymbol{Y_{t+l-j}}$ para $l \leq j$. Da representação MA de ordem infinita (6.69) tem-se que a previsão associada à expressão (6.77) pode também ser determinada a partir da expressão

$$\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}(\boldsymbol{l}) = \sum_{\boldsymbol{j=l}}^{\infty} \boldsymbol{\Psi_j^*}\, \boldsymbol{X_{t+l-j}} + \sum_{\boldsymbol{j=l}}^{\infty} \boldsymbol{\Psi_j}\, \boldsymbol{\varepsilon_{t+l-j}}$$

pelo que o erro de previsão é

$$\boldsymbol{e_t(l)} = \boldsymbol{Y_{t+l}} - \widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}(\boldsymbol{l}) = \sum_{\boldsymbol{j=0}}^{\boldsymbol{l-1}} \boldsymbol{\Psi_j}\, \boldsymbol{\varepsilon_{t+l-j}}$$

tem-se então que a matriz MSE dos erros de previsão "condicional" l períodos à frente de $\tilde{Y}_t(l)$ será dada por

$$\boldsymbol{\Sigma(l)} = \boldsymbol{E}[\boldsymbol{e_t(l)e_t'(l)}] = \sum_{\boldsymbol{j=0}}^{\boldsymbol{l-1}} \boldsymbol{\Psi_j}\, \boldsymbol{\Sigma_\varepsilon} \boldsymbol{\Psi_j'} \tag{6.78}$$

Comparando as duas expressões, (6.76) e (6.78), verifica-se que $\boldsymbol{\Sigma(l)}$ dado por esta ultima expressão é igual ao segundo termo da expressão (6.76). Este termo corresponde aos erros de previsão devido à componente MA do modelo ARMAX, ao passo que primeiro termo do segundo membros de (6.76) correspondente ao erro na previsão de valores futuros da variável exógena.

6.2.3 Controlo óptimo

Em sistemas económicos, físicos, biológicos, processos industriais etc., alguma variáveis exógenas, nomeadamente as chamadas variáveis manipuláveis, podem ser utilizadas para manter ou conduzir os sistemas aos objectivos desejados ou perto deles. Sabendo-se quais os valores objectivo das variáveis endógenas, pode-se então ajustar as variáveis manipuladas para que as previsões alguns passos à frente, obtidas com recurso ao modelo do processo, estejam de acordo com os objectivos definidos.

Normalmente não será de todo possível atingir os objectivos para todas as variáveis endógenas, até porque normalmente estão correlacionadas entre si, ou porque as variáveis exógenas tem uma gama de operação limitada. Consequentemente haverá a necessidade de se recorrer a uma função de custo para se conciliar os interesses envolvidos e a disponibilidade de manipulação.

6.2.3.1 Controlo óptimo em modelos ARMAX

Os modelos ARMAX poderão assim revelar-se como uma ferramenta bastante útil para optimizar o controlo dos sistemas. Especificamente, assume-se que algumas variáveis exógenas X_t são manipuláveis (as variáveis exógenas conhecidas mas não manipuladas, nesta secção, serão referidas por $\boldsymbol{W_t}$). Obtida a informação ao longo do tempo, pode-se determinar o "valor óptimo" $\boldsymbol{X^*_{t+1}}$ das variáveis manipuláveis no instante $t+1$ de forma a que as variáveis endógenas $\boldsymbol{Y_{t+1}}$ estejam próximas dos valores pretendidos $\boldsymbol{Y^*_{t+1}}$ (Target) no instante $t+1$. Pode também pretender-se que o valor de $\boldsymbol{X^0_{t+1}}$ oscile pouco em relação a um determinado valor alvo pretendido para as variáveis exógenas, ou que não varie excessivamente em relação ao nível anterior. Desta forma, o problema de controlo óptimo um passo à frente pode ser formulado como a determinação do valor $\boldsymbol{X^*_{t+1}}$ que minimize a função de custo, ou índice de performance J

$$\boldsymbol{J = E_t\left[\left(Y_{t+1} - Y^0_{t+1}\right)' Q_{t+1}\left(Y_{t+1} - Y^0_{t+1}\right) + \left(X_{t+1} - X^0_{t+1}\right)' P_{t+1}\left(X_{t+1} - X^0_{t+1}\right)\right]} \quad (6.79)$$

onde $\boldsymbol{E_t}[\,]$ significa o valor esperado condicionado à informação obtida até ao instante t, e $\boldsymbol{Q_{t+1}}$ e $'\boldsymbol{P_{t+1}}$ são matrizes não negativas definidas que aplicam as ponderações relativas aos desvios $\boldsymbol{Y_{t+1} - Y^0_{t+1}}$ e $\boldsymbol{X_{t+1} - X^0_{t+1}}$ em relação aos valores alvo das respostas e das variáveis manipuladas, respectivamente.

Para determinar a leis de controlo, é conveniente separar os dados conhecidos no instante t dos dados desconhecidos. Pelo que o modelo ARMAX (6.68) pode ser simplesmente como

$$\boldsymbol{Y_{t+1} = B' X_{t-1} + b_t + c_{t+1} + e_{t+1}} \quad (6.80)$$

Onde $\boldsymbol{B' = \Psi_0}$, $b_t = \sum_{j=1}^{p} \Phi_j\, Y_{t+1-j} = \sum_{j=1}^{s} \Theta^*_j\, X_{t+1-j} - \sum_{j=1}^{q} \Theta_j\, \varepsilon_{t+1-j}$ são as componentes de Y_{t+1} que contém as variáveis já determinadas no instante t e c_{t+1} é um termo adicional que contém alguns aspectos de futuras variáveis exógenas não manipuladas, cujos valores se assumem conhecidos.

Substituindo a expressão anterior na função de custo, tem-se, após substituição de e_{t+1} do respectivo valor esperado

$$\boldsymbol{J = E_t\Big[\left(B^{'}X_{t-1} + b_t + c_{t+1} - Y^0_{t+1}\right)^{'} Q_{t+1}\left(B^{'}X_{t-1} + b_t + c_{t+1} - Y^0_{t+1}\right) + \left(X_{t+1} - X^0_{t+1}\right)^{'} P_{t+1}\left(X_{t+1} - X^0_{t+1}\right)\Big]} \quad (6.81)$$

Para minimizar a função de custo, procede-se à sua diferenciação em ordem a X_{t+1} e iguala-se a zero, obtendo-se (Reinsel, 1997)

$$\boldsymbol{\frac{dJ}{dX_{t+1}} = 2\left\{BQ_{t+1}\left(B^{'}X_{t-1} + b_t + c_{t+1} - Y^0_{t+1}\right) + P_{t+1}\left(X_{t+1} - X^0_{t+1}\right)\right\} = 0}$$

Que resulta no valor óptimo das variáveis de controlo

$$\boldsymbol{X^*_{t+1} = \left(BQ_{t+1}B^{'} + P_{t+1}\right)^{-1}\left[BQ_{t+1}\left(Y^0_{t+1} - b_t - c_{t+1}\right) + P_{t+1}X^0_{t+1}\right]} \quad (6.82)$$

Este resultado pode ser generalizado para os casos em que a resposta do sistema não é instantânea como no caso estudado, ou seja, quando há um atraso $b > 0$ na resposta Y_t em relação à acção de controlo $\boldsymbol{X_t}$. Isto é, assume-se que o modelo ARMAX (6.68) pode ser dado por

$$\boldsymbol{Y_t = \sum_{j=1}^{p} \Phi_j Y_{t-j} + B^{'}X_{t-b} + \sum_{j=1}^{s} \Theta^*_j + \varepsilon_t - \sum_{j=1}^{q} \Theta_j \varepsilon_{t-j}}$$

O valor de $\boldsymbol{X_{t+1}}$ não influencia a saída antes do instante $t + b + 1$, de modo a que no instante t, o problema de controlo é determinar o valor de $\boldsymbol{X^*_{t+1}}$ de X_{t+1} que minimize

$$\boldsymbol{J = E_t\left[\left(Y_{t+b+1} - Y^0_{t+b+1}\right)'Q_{t+b+1}\left(Y_{t+1} - Y^0_{t+1}\right) + \left(X_{t+1} - X^0_{t+1}\right)'P_{t+1}\left(X_{t+1} - X^0_{t+1}\right)\right]}$$

Neste caso, considera-se a previsão $(b + 1)$ períodos à frente de $\boldsymbol{Y_{t+b+1}}$ condicional à informação existente no instante t. Pelo que, seguindo o mesmo raciocínio descrito atrás, a equação de controlo óptimo é dada por (Reinsel, 1997)

$$\boldsymbol{X^*_{t+1} = \left(BQ_{t+b+1}B^{'} + P_{t+1}\right)^{-1}\left[BQ_{t+b+1}\left(Y^0_{t+b+1} - b_t - c_{t+1}\right) + P_{t+1}X^0_{t+1}\right]} \quad (6.83)$$

A expressão (6.79), com ligeiras alterações, mas sem perca de objectividade poderá ser escrita numa forma mais intuitiva

$$\boldsymbol{J = \left[\left(\hat{Y}_{t+1|t} - T\right)'Q\left(\hat{Y}_{t+1|t} - T\right) + \left(X_t - X_{t-1}\right)P\left(X_t - X_{t-1}\right)\right]} \quad (6.84)$$

Olhando para as duas expressões, verifica-se que no último caso, as ponderações são constantes (invariantes no tempo) e que o processo pressupõe-se instantaneamente não causal. Em termos de implementação, não existem grandes dificuldades em passar de uma expressão para a outra. Contudo, olhando para a última expressão e para a expressão (5.15), verifica-se que as semelhanças são evidentes, pelo que esta expressão é conhecida como a versão multivariada do índice de performance de Clarke e Gawthrop (Del Castillo, An adaptative run-to-run optimizing controller for linear and nonlinear semicndutor processes, 1998).

Decompondo as variáveis exógenas nos três tipos admissíveis, denominando $X1$ como as variáveis exógenas cujos níveis são conhecidos mas que não são passíveis de alteração ou manipulação; X2 como as variáveis exógenas que se podem manipular com a finalidade de controlar as respostas do sistema; e como X3 as variáveis exógenas cujos valores futuros tem que ser previstos, ou seja, variáveis exógenas com dinâmica; tem-se então que X pode ser definido pelo vector $X = (X1^{'}, X2^{'}, X3')^{'}$. Com estes pressupostos, um diagrama de blocos de uma possível implementação do controlador Clarke e Gawthrop multivariado pode dado pela Figura 4-1

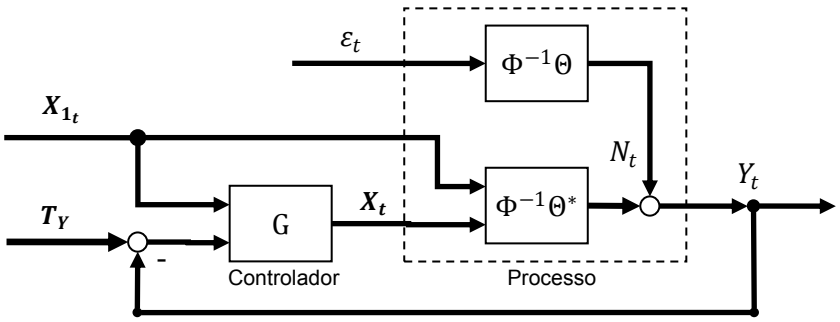


Figura 6-3 Possível implementação de um sistema de controlo baseado no modelo ARMAX multivariado

6.2.4 Especificação e validação do modelo ARMAX

Definindo o vector das variáveis exógenas $\boldsymbol{X} = (X_1^{'}, X_2^{'}, X_3^{'})^{'}$ como descrito acima, o modelo ARMAX pode ser representado numa forma especial, e particularmente útil, de um vector ARMA para o processo combinado $\boldsymbol{Z_t} = (\boldsymbol{X_t^{'}}, \boldsymbol{Y_t^{'}}) = (X_{1t}^{'}, X_{2t}^{'}, X_{3t}^{'}, \boldsymbol{Y_t^{'}})$. Ou seja, o vector $\boldsymbol{Z_t}$ tem a seguinte estrutura

$$\boldsymbol{Z_t} = \left[\boldsymbol{X_{1_{1,t}}^{'}}, \ \dots, \boldsymbol{X_{1_{r_1,t}}^{'}}, \ \boldsymbol{X_{2_{1,t}}^{'}}, \ \dots, \ \boldsymbol{X_{2_{r_2,t}}^{'}}, \ \boldsymbol{X_{3_{1,t}}^{'}}, \ \dots, \ \boldsymbol{X_{3_{r_3,t}}^{'}}, \boldsymbol{Y_{1t}^{'}}, \ \dots, \boldsymbol{Y_{kt}^{'}}\right]$$

em que $\boldsymbol{r_1} + \boldsymbol{r_2} + \boldsymbol{r_3} = r$ é o numero total de variáveis exógenas.

Admitindo que X_3 é representado por um processo ARMA dado pela expressão (6.70), ou mesmo por um sistema ARMAX em que as variáveis X_1 e X_2 fazem o papel de variáveis exógenas com a restrição de não apresentarem dinâmica, ou seja, neste caso seria um sistema ARMAX apenas com variáveis determinísticas ou previamente conhecidas. Ou seja, X_3 dado por

$$\boldsymbol{X_{3_t}} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \mathbf{A}_j \boldsymbol{X_{3_{t-j}}} = \sum_{\boldsymbol{j=0}}^{\boldsymbol{s}} \left[\mathbf{C}_{1_j}^{*} \boldsymbol{X_{1_{t-j}}} + \mathbf{C}_{2_j}^{*} \boldsymbol{X_{2_{t-j}}}\right] + \boldsymbol{a_{3_t}} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{q}} \mathbf{C}_j \boldsymbol{a_{3_{t-j}}}$$

Admitindo ainda, por questões de simplificação mas sem perda de generalidade, que o sistema (6.68) é instantaneamente não causal, ou seja $\boldsymbol{\Theta}_0^* = 0$. Nestas condições o sistema ARMAX pode ser representado pelo modelo ARMA para a variável Z dado por

$$\boldsymbol{Z_t} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\bar{\Phi}_j}\, \boldsymbol{Z_{t-j}} = \boldsymbol{u_t} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{q}} \boldsymbol{\bar{\Theta}_j}\, \boldsymbol{u_{t-j}} \tag{6.85}$$

onde

$$\boldsymbol{u_t} = \begin{bmatrix} \boldsymbol{a_{1_t}} \\ \boldsymbol{a_{2_t}} \\ \boldsymbol{a_{3_t}} \\ \boldsymbol{\varepsilon_t} \end{bmatrix}$$

$$\boldsymbol{\bar{\Phi}_j} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{C}_{1_j}^{*} & \mathbf{C}_{2_j}^{*} & \mathbf{A}_j & \mathbf{0} \\ \boldsymbol{\Theta}_{1_j}^{*} & \boldsymbol{\Theta}_{2_j}^{*} & \boldsymbol{\Theta}_{3_j}^{*} & \boldsymbol{\Phi}_j \end{bmatrix}$$

$$\boldsymbol{\bar{\Theta}_j} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{C}_j & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\Theta}_j \end{bmatrix}$$

Deste modo, os métodos para especificação do modelo ARMAX são bastantes similares aos descritos previamente para o modelo ARMA, com a imposição das devidas restrições.

A análise dos resíduos $\hat{\varepsilon}_t$ é útil para validar a adequação do modelo especificado. A examinação da autocorrelação, em particular, é bastante importante para validar o pressuposto do modelo de que $\hat{\varepsilon}_t$ é um processo ruído branco. Outra verificação que é importante fazer é examinar a correlação cruzada entre valores do ruído do modelo ARMAX $\hat{\varepsilon}_t$ e os valores do ruído do modelo ARMA da variável exógena $\hat{a}_t$. A existência substancial de correlação cruzada entre $\hat{\varepsilon}_t$ e valores desfasados (*lagged*) de $\hat{a}_{t-j}$ $(j \geq 0)$, poderão indiciar alguma inadequação na especificação da estrutura de desfasamentos para a variável exógena no modelo ARMAX, que deverá ser corrigida.

7 Combinação SPC e EPC

Existe alguma confusão sobre o ajustamento ou regulação do processo e o papel importante que desempenha na redução da variabilidade. Por exemplo, a metodologia das cartas de controlo não é o melhor método para reduzir a variabilidade na vizinhança dos valores de referência. Na indústria de processo e química, a utilização de técnicas, tais como simples regras de controlo integral têm sido efectivamente utilizadas para esse fim.

A monitorização do processo e o ajustamento do processo são duas abordagens complementares para controlo do processo. Estes procedimentos partilham o objectivo comum que é a redução da variabilidade do processo (Montgomery, 2001); (Box, Jenkins, & Reinsel, 2008). A monitorização do processo, que é parte do controlo estatístico do processo (SPC), foca os seus objectivos em identificar causas especiais e então eliminar as fontes das causas, conseguindo assim reduzir a variabilidade do processo e melhorar o seu desempenho. Paralelamente, o ajustamento do processo (EPC) foca o seu objectivo em manter a saída do processo próxima dos valores de referência recorrendo à manipulação de das variáveis de entrada.

Em geral, a teoria de controlo de engenharia está baseada na ideia de que, verificando-se os seguintes pressupostos:

1. É possível prever a próxima observação do processo;
2. Existem outras variáveis que podem ser manipuladas de modo a afectar a saída do processo;
3. Conhece-se o efeito das variáveis manipuladas tal que se consegue determinar qual a acção de controlo a aplicar;

Então pode-se fazer o ajustamento nas variáveis manipuladas no instante t que seja mais provável para produzir uma saída no processo coincidente com a referência no instante $t+1$. Claramente que esta ideia requer um conhecimento da relação entre a saída do processo e as variáveis manipuladas, bem como compreensão das dinâmicas do processo. É conveniente ainda que exista alguma facilidade em manipular as variáveis de entrada. De facto, se o custo de efectuar as acções de controlo é desprezível, então tem-se que a variabilidade do output do processo é minimizada procedendo ao ajustamento em todos os instantes. De notar que aqui existe uma diferença significativa em relação ao SPC, onde as acções de controlo apenas são desencadeadas quando existe uma evidência estatística de que o processo está fora de controlo. Essa evidência é normalmente um ponto fora dos limites da carta de controlo.

Existem muitos processos onde alguns tipos de esquemas de controlo por feedback serão preferíveis às cartas de controlo. Por outro lado o EPC por si só, não tenta identificar causas especiais que possam ter impacto no processo, se bem que o simples facto dos actuadores saturarem poderá por si só desencadear o alerta para uma possível causa especial. A eliminação das causas especiais podem resultar numa melhoria significativa dos processo. Em principio, o que a metodologia EPC faz é reagir aos desvios do output em relação aos valores de referência, pelo que não efectuará qualquer esforço no sentido de remover

qualquer causa especial. Consequentemente, alguns dos processos que utilizam apenas o controlo por feedback, podem ser significativamente melhorados com a implementação conjunta de cartas de controlo para a monitorização estatística do processo. Este tipo de implementação é por vezes referido como SPC algorítmico (*algoritmic SPC ou ASPC - Algoritmic Statistical Process Contro*l) (Vander Weil S. T., 1992).

O controlo estatístico deve ser aplicado às variáveis de entrada e às variáveis de saída. Pontos fora dos limites de controlo devem identificar períodos onde os erros de controlo são grandes ou onde as acções de controlo são elevadas devido a compensações excessivas devido a causas especiais, que deste modo podem ser detectadas. Esses períodos deverão provavelmente constituir boas oportunidades para se proceder à procura de causas especiais. De notar que o EPC tem como pressuposto a existência de um modelo interno em que os resíduos (desvios em relação à referência) seguem uma distribuição bem definida - $\{\varepsilon_t\} \rightsquigarrow N(0, \Sigma_\varepsilon)$, pelo que qualquer alteração desta distribuição poderá significar uma alteração do processo em relação ao modelo que se considerava adequado. A combinação do SPC/EPC poderá ser implementada de acordo com o diagrama de blocos da Figura 7-1 (Montgomery D. C., 2001)

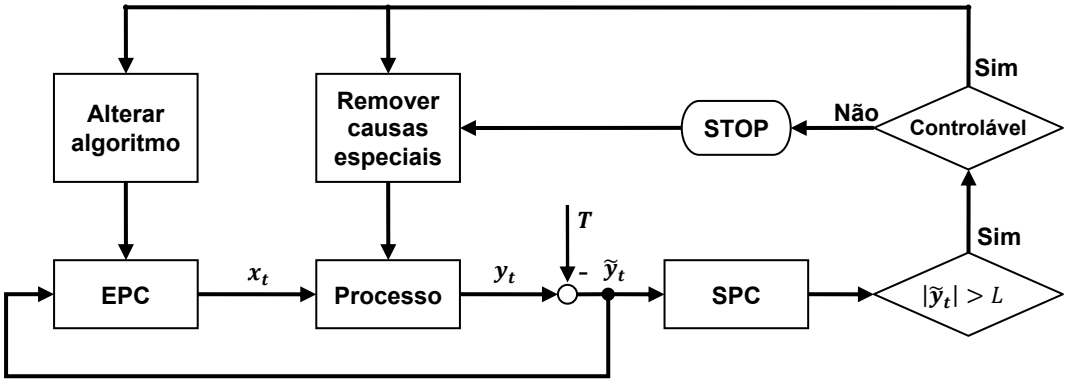


Figura 7-1 Integração EPC/SPC baseado em (Montgomery D. C., 2001)

7.1 Controlo económico

A engenharia controlo do processo (EPC) tem-se desenvolvido largamente na industria química e de processo, particularmente na industria química onde a compensação continua é fundamental para cancelar os efeitos dos distúrbios não controlados. Em contraste, o desenvolvimento do controlo estatístico do processo (SPC) está mais ligado à indústria de produção discreta onde é possível obter-se um produto quase uniforme devido à estandardização dos procedimentos e matérias-primas e pela aplicação de um sistema de monitorização para a detecção e remoção de fontes de problemas.

Enquanto que as duas abordagens reflectem características das suas diferentes origens, em quase todas as industrias é necessário algum conhecimento de SPC e EPC para se conseguirem as melhores práticas de produção. Isto torna-se mais verdade actualmente em que produtos e processos desconhecidos são continuamente desenvolvidos. Por exemplo, a industria electrónica utiliza métodos híbridos de produção em que num estágio lida-se com processos químicos e noutros com produção discreta.

Quando se utiliza controlo contínuo por realimentação, normalmente assume-se a suposição táctica de que os custos envolvidos são apenas os custos de não conformidade (produto fora dos limites de especificação). No entanto, particularmente na produção discreta, tem que se contabilizar os custos adicionais inerentes aos procedimentos de amostragem e ajustamento.

7.1.1 Controlo de custo mínimo com custos fixos de ajustamento e monitorização

Num cenário em que se assume que apenas os custos de controlo a considerar são os custos relacionados com os desvios quadráticos em relação à referência (custos *off target*), um controlador MMSE sem restrições implicará a minimização de todos os custos. Suponha-se, no entanto, que existem outros custos quadráticos adicionais associados por exemplo à amplitude dos ajustamentos das variáveis de entrada x_t, e que α é uma constante de proporcionalidade que relaciona os custos de *off target* com os custos de ajustamento, então o custo global de controlo será proporcional a $(\sigma_\varepsilon^2 + \alpha\sigma_x^2)$, pelo que a minimização desta quantidade resultará no custo global mínimo. Em vez da formulação deste cenário, que pode conduzir a valores teóricos ideais mas não satisfatórios na prática, pode-se formular outro cenário alternativo: o que é que constitui uma redução satisfatória σ_x^2 em troca de um aceitável incremento de σ_ε^2. O mesmo raciocínio poderá ser aplicado a sistemas com custos de ajustamento e monitorização.

7.1.1.1 Ajustamento condicionado para custos de ajustamento fixos com modelo de distúrbios IMA(*1,1*)

A maioria dos processos na indústria de produção discreta são considerados processos responsivos, ou seja, o ajustamento efectuado tem efeito total imediato. Mas normalmente estes ajustamentos tem custos fixos inerentes, como por exemplo, ter que parar uma máquina e mudar uma ferramenta.

Para estimar os custos de *off target*, pode recorrer-se uma função quadrática fácil de manipular matematicamente

$$\boldsymbol{c = k_t(y' - T)^2} \tag{7.1}$$

Onde $\boldsymbol{c}$ é o custo de *off target* e $\boldsymbol{k_t}$ é uma constante. Para esta função de custo quadrática, a minimização do custo é equivalente a minimizar o MMSE. A constante $\boldsymbol{k_t}$ é o custo do processo estar a uma unidade de distância do valor de referência para um período de tempo. Em geral, torna-se mais conveniente trabalhar, não com $\boldsymbol{k_t}$, mas com a constante normalizada $\boldsymbol{C_T = k_t\sigma_\varepsilon^2}$**.** Neste caso, o custo *off target* pode ser dado pela expressão

$$\boldsymbol{c = C_T\left(\frac{y_t' - T}{\sigma_\varepsilon}\right)^2} \tag{7.2}$$

e C_T é o custo médio de *off target* para um período por uma quantidade de σ_ε. No pressuposto de que os distúrbios podem ser modelados por um modelo IMA(*1,1*), σ_ε é o desvio padrão que o processo exibe quando não existe limite para se proceder ao

ajustamento, ou seja, quando o valor de L (distância que limita a zona morta $[T-L, T+L]$) é igual a zero $(L=0)$ (ver Figura 7-2)

Se a função de custo for quadrática com um mínimo em T e assumindo-se σ_ε conhecido, para se determinar a constante C_T, apenas se necessita de um ponto adicional da curva de custo. (Taguchi, 1987) diz que para se obter um ponto adicional basta conhecer o desvio a partir do qual resulta a rejeição de todo o material produzido durante um período de amostragem com um custo de c_0, ou seja

$$\boldsymbol{c_0 = c_T\left(\frac{y_0^{'} - T}{\sigma_\varepsilon}\right)^2}$$

e consequentemente

$$\boldsymbol{c_T = \frac{c_0 \sigma_\varepsilon^2}{(y_0^{'} - T)^2}}$$

Para se construir uma carta de ajustamento limitado, é necessário conhecer-se o valor que L tal que as linhas limites $T \pm L$ resultará no menor custo global (Figura 7-2). (Box & Luceño, 1997) apresenta uma tabela com valores de $L/\lambda\sigma_\varepsilon$, do intervalo médio entre ajustamentos (AAI) e de q para vários valores de $R_a = (C_A/C_T)/\lambda^2$, em que $\lambda = 1-\theta$ é a constante de amortecimento. O erro quadrático médio do processo ajustado é então dado por $\boldsymbol{\sigma_\varepsilon^2}(1+\lambda^2 q)$ (Box & Luceño, 1997), ou seja, o incremento no desvio padrão (ISD) será dado por $ISD = (1+\lambda^2 q)^{1/2}$.

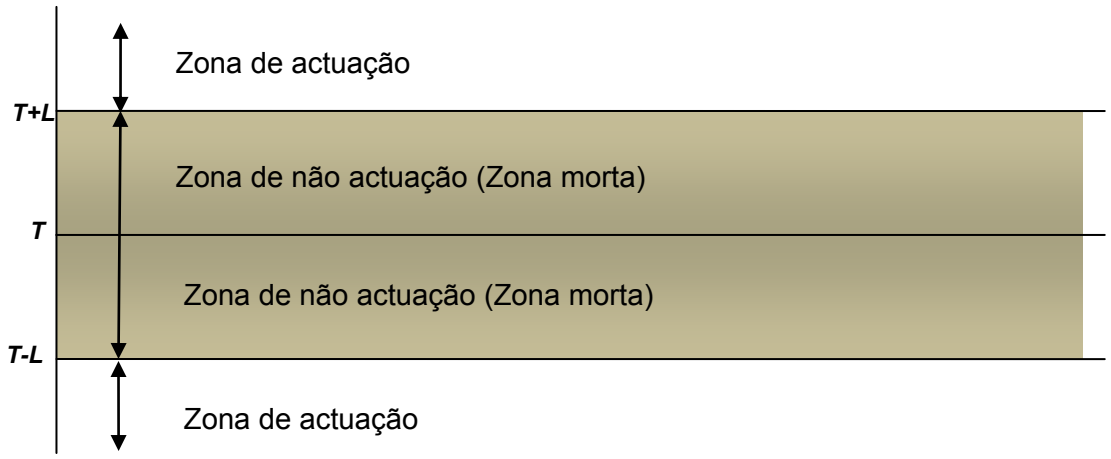


Figura 7-2 Limites de zona morta para cenários de controlo de custo mínimo

7.1.1.2 Inclusão dos custos de amostragem

Outros dos possíveis custos envolvidos no controlo do processo são os custos de amostragem, pelo que será de conveniente incluir estes custos na determinação do custo global mínimo. A questão passa então a ser qual a frequência de amostragem que reduz os custos globais.

Suponha-se que o no processo em estudo, ou que durante o período de identificação do processo, as observações foram recolhidas com um intervalo de amostragem específico, ao

qual se passou a chamar de intervalo unitário, suponha-se ainda que exista a pretensão de reduzir a frequência de amostragem por razões económicas ou outras. Tal como no caso anterior, admite-se que o sistema é responsivo (sem dinâmica) e que os distúrbios são aproximadamente modelados por uma série IMA(1,1), que o custo *off target* (C_T) proporcional ao erro quadrático médio, e que existe um custo de ajustamento fixo (C_A). Introduz-se agora um terceiro custo (C_S), o custo de amostragem.

Devido à dificuldade de determinar os três custos envolvidos, C_T, C_A e C_S, (Box & Luceño, 1997) baseado nas seguintes condições

1. O intervalo S é expresso em função do intervalo unitário;
2. O intervalo entre ajustamentos (AAI) é também expresso em termos do intervalo unitário;
3. A resultante percentagem de incremento no desvio padrão (ISD) é expressa em termos de σ_ε;

fornece uma série de ábacos (Box & Luceño, 1997, pp. 221-222) para valores de $\lambda = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6$ que permitem determinar o melhor valor de L para o cenário em causa. Para cada valor de λ, os ábacos dão o valor do incremento do desvio padrão (ISD) e o correspondente intervalo médio entre ajustamentos (AAI) para vários valores de S e L/σ_ε.

Suponha-se que uma série temporal (denomine-se por serie 1) , modelada por uma IMA com parâmetro de não estacionaridade $\lambda = \lambda_1$ e gerada por um processo ruído branco $\{\varepsilon_t\}$ com desvio padrão $\sigma_\varepsilon = \sigma_{\varepsilon 1}$, foi amostrada com a periodicidade de S intervalos unitários (ou seja de S em S intervalos da série original) gerando assim uma serie *S*. Então pode-se demonstrar (Box, Jenkins, & Reinsel, 2008) que a série amostrada (serie S) pode também ser modelada por um IMA com parâmetro λ_S também gerado por um processo ruído branco com desvio padrão $\sigma_{\varepsilon S}$, e as relações entre os parâmetros da série original (serie 1) e a série amostrada (série S) são

$$\frac{S\lambda_1^2}{1-\lambda_1} = \frac{\lambda_S^2}{1-\lambda_S}$$

e

$$\sigma_{\varepsilon S} = \sigma_{\varepsilon 1}\sqrt{\frac{1-\lambda_1}{1-\lambda_S}} = \sigma_{\varepsilon 1}\sqrt{\frac{\theta_1}{\theta_S}}$$

Para valores pretendidos de $\boldsymbol{\lambda}$, S e $L/\boldsymbol{\sigma_\varepsilon}$, os ábacos fornecem valores de AAI e ISD subjacentes aos pressupostos de que os distúrbios seguem um modelo IMA, o sistema é responsivo e a função de custo é quadrática. Nesta condições, para um dado $\boldsymbol{\lambda}$, pode-se determinar os valores de S e $L/\boldsymbol{\sigma_\varepsilon}$ necessários para produzir qualquer combinação de valores de ISD e AAI.

Quando os custos podem ser estimados, os esquemas de ajustamento por feedback podem ser caracterizados directamente em termos dos custos normalizados C_T, C_A e C_S, em vez dos indicadores AAI e ISD, recorrendo a ábacos desenvolvidos por (Box G. E., 1992) e reproduzidos em (Box & Luceño, 1997, pp. 228-229). Através destes ábacos é possível determinar aproximadamente o dos limites normalizados de zona morta $l = L/\lambda\boldsymbol{\sigma_\varepsilon}$ e o intervalo de amostragem S em função dos valores de $\boldsymbol{\lambda}$, $R_A = (C_A/C_T)/\lambda^2$ e $R_S = (C_S/C_T)/\lambda^2$.

Cálculo numérico
Os ábacos, apesar de fornecerem normalmente aproximações adequadas, não são particularmente fáceis de ler. Alternativamente, as quantidades L e S que minimizam os custos esperados, podem ser determinadas a partir de um algoritmo de cálculo numérico desenvolvido por (Luceño, Gonzalez, & Puig-Pey, 1996) e que foi implementado sobre a plataforma Visual C++ no âmbito deste trabalho (Anexos 1 e 2).

7.2 Monitorização dos parâmetros e estratégias de ajustamento por realimentação

Outra estratégia de integração da engenharia de controlo com o controlo estatístico passa por monitorizar os próprios parâmetros do modelo e consequentemente do controlador, através de cartas de controlo.

Uma estratégia de aplicação de controlo estatístico a processos com dados correlacionados, passa por filtrar os dados até que os resíduos exibam um comportamento modelado por um processo ruído branco (secção 6.1.4.3) (Figura 6-1), ou seja, os resíduos resultantes devem ser independentes e identicamente distribuídos (iid).

O controlo estatístico baseado nos resíduos é dos métodos estatísticos mais largamente investigados. Como tem sido referido ao longo deste texto, a ideia base por detrás das cartas de controlo dos resíduos é ajustar um modelo de série temporal que represente a autocorrelação. Se o modelo é adequado, os resíduos são aproximadamente estatisticamente independentes, e então cartas de Shewhart, CUSUM ou EWMA, podem ser aplicadas aos resíduos.

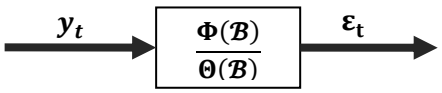


Figura 7-3 Processo representado por um modelo ARMA

Suponha-se então que se tem um processo representado por um modelo ARMA(*p,q*) com esquematizado na Figura 7-3, dado por

$$x_t = \frac{\Theta(B)}{\Phi(B)}\varepsilon_t \tag{7.3}$$

Onde x_t são os dados observados e ε_t são variáveis normais independentes e identicamente distribuídas com média zero e variância σ_ε^2. O modelo **(7.3**) denomina-se por modelo nulo. Suponha-se que existe uma mudança determinística, que será referida como uma "falha", no processo no instante τ. Então, a partir do instante τ, os dados do processo a serem observados poderão ser representados pelo modelo

$$y_t = x_t + \mu f_{t-\tau}$$

onde f_t indica a natureza da falha e μ a amplitude. Este modelo denomina-se por modelo com discrepância. Por exemplo, a função f_t para uma alteração em degrau pode ser definida por

$$\boldsymbol{y_t} = \begin{cases} 1 & se\ \ t \geq 0 \\ 0 & se\ \ t < 0 \end{cases}$$

Assumindo-se que o modelo **(7.3**) é invertível, então os resíduos podem ser obtidos filtrando $\boldsymbol{y_t}$ com o filtro inverso $\boldsymbol{\Phi(\mathcal{B})/\Theta(\mathcal{B})}$, ou seja

$$\boldsymbol{e_t} = \frac{\boldsymbol{\Phi(\mathcal{B})}}{\boldsymbol{\Theta(\mathcal{B})}}\boldsymbol{y_t} = \frac{\boldsymbol{\Phi(\mathcal{B})}}{\boldsymbol{\Theta(\mathcal{B})}}(\boldsymbol{x_t} + \mu \boldsymbol{f_{t-\tau}})$$

$$\boldsymbol{e_t} = \underbrace{\frac{\boldsymbol{\Phi(\mathcal{B})}}{\boldsymbol{\Theta(\mathcal{B})}}\boldsymbol{x_t}}_{\boldsymbol{\varepsilon_t}} + \mu \underbrace{\frac{\boldsymbol{\Phi(\mathcal{B})}}{\boldsymbol{\Theta(\mathcal{B})}}\boldsymbol{f_{t-\tau}}}_{\tilde{\boldsymbol{f}}_{\boldsymbol{t-\tau}}} \tag{7.4}$$

ou seja

$$\boldsymbol{e_t} = \boldsymbol{\varepsilon_t} + \mu \tilde{\boldsymbol{f}}_{\boldsymbol{t-\tau}}$$

onde $\tilde{\boldsymbol{f}}_{\boldsymbol{t-\tau}} = \boldsymbol{\Phi(\mathcal{B})/\Theta(\mathcal{B})}$ denomina-se como sendo assinatura da falha. Deste modo, os resíduos estão não correlacionados com uma média variante com o tempo e variância σ_ε^2. O valor de $\tilde{\boldsymbol{f}}_{\boldsymbol{t}}$ depende do modelo ARMA e consequentemente da estrutura de autocorrelação dos dados.

A informação contida na dinâmica da assinatura da falha pode ser útil na detecção da própria falha. As cartas tradicionais não fazem uso desta informação, mas a carta de "resultados" acumulados CuScore (*Cumulative Score*) podem ter essa informação em conta. O objectivo do teste CuScore é detectar alterações nos parâmetros de um modelo estatístico, que é uma aplicação natural das cartas de controlo.

A estatística CuScore unilateral é dada por

$$\boldsymbol{Q_t} = \boldsymbol{max}\{\boldsymbol{Q_{t-1}} + \boldsymbol{r_t}(\boldsymbol{e_t} - \boldsymbol{k}), \boldsymbol{0}\} \quad \boldsymbol{t} = \boldsymbol{1, 2, \ldots} \tag{7.5}$$

onde $\boldsymbol{r_t}$ e k são respectivamente o detector e o valor de referência. Se $\boldsymbol{Q_t}$ excede o intervalo de decisão, h, então conclui-se que ocorreu uma falha no processo.

7.2.1.1 Estatística Cuscore

A estatística Cuscore é um procedimento projectado para detectar alterações nos parâmetros de um modelo estatístico. Considere-se um modelo estatístico escrito na forma

$$\boldsymbol{e_t} = \boldsymbol{f}(\boldsymbol{y_t}, \theta) \tag{7.6}$$

onde $\boldsymbol{y_t}$ é a observação e θ é um qualquer parâmetro do modelo. A estatística Cuscore associada com o valor $\theta = \theta_0$ é

$$\boldsymbol{Q_t} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{t}} \boldsymbol{\varepsilon_{i0} r_i} \tag{7.7}$$

onde os valores $\boldsymbol{\varepsilon_{i0}}$ (modelo nulo) obtêm-se fazendo $\theta = \theta_0$ na equação **(7.6**) e em que o detector $\boldsymbol{r_i}$ é dado por

$$\boldsymbol{r_i} = -\frac{\boldsymbol{\partial \varepsilon_i}}{\boldsymbol{\partial \theta}}\Big|_{\theta=\theta_0}$$

Consequentemente, a estatística Cuscore pode ser vista como uma soma acumulada de produtos de resíduos do modelo nulo $\boldsymbol{\varepsilon_{i0}}$ com um detector apropriado $\boldsymbol{r_i}$.

Do modelo com discrepância (7.4) tem-se que $\boldsymbol{\varepsilon_t} = \boldsymbol{e_t} - \mu\tilde{\boldsymbol{f}}_{\boldsymbol{t-\tau}}$. Quando não existe nenhuma alteração, assume-se que $\mu = 0$, pelo que o modelo nulo será dado por $\boldsymbol{\varepsilon_{t0}} = \boldsymbol{e_t}$. Quando existe uma alteração, os resíduos passam a ter média $\mu\tilde{\boldsymbol{f}}_{\boldsymbol{t-\tau}}$, pelo que deixam de ser ruído branco. Detectar a alteração é equivalente a detectar a alteração de μ em relação à sua média anterior (que era zero), ou mais precisamente, procurando $\tilde{\boldsymbol{f}}_{\boldsymbol{t-\tau}}$ nos resíduos. O detector para detectar a alteração de μ em relação ao valor original que era zero é

$$\boldsymbol{r_i} = -\frac{\boldsymbol{\partial \varepsilon_i}}{\boldsymbol{\partial \mu}}\Big|_{\mu=0} = \tilde{\boldsymbol{f}}_{\boldsymbol{i-\tau}}$$

O valor de $\boldsymbol{\varepsilon_{i0}}$ é

$$\boldsymbol{\varepsilon_{i0}} = \boldsymbol{\varepsilon_i}|_{\mu=0} = \boldsymbol{e_i}$$

Deste modo, a estatística Cuscore para monitorizar a média de um processo normalmente distribuído e autocorrelacionado é

$$\boldsymbol{Q_t} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{t}} \boldsymbol{\varepsilon_{i0} r_i} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{t}} \boldsymbol{e_i r_i} \tag{7.8}$$

Onde $\boldsymbol{r_i} = \tilde{\boldsymbol{f}}_{\boldsymbol{i-\tau}}$.

Para implementação computacional, é conveniente reescrever esta estatística na forma recursiva

$$\boldsymbol{Q_t} = \boldsymbol{Q_{t-1}} + \boldsymbol{e_t r_t}$$

Esta estatística, contudo, não deve ser utilizada directamente. A razão é que a maioria dos sistemas começam com um longo período em que o processo não está controlado, durante o qual $\boldsymbol{Q_t}$ vagueia por zonas fora de controlo. Quando uma falha ocorre, $\boldsymbol{Q_t}$ pode não se situar exactamente numa situação de controlo. Pelo que deve ser substituída pela estatística Cuscore unilateral (Shu, Apley, & Tsung, 2002)

$$\boldsymbol{Q_t} = \boldsymbol{max}\{(\boldsymbol{Q_{t-1}} + \boldsymbol{e_t r_t}), \boldsymbol{0}\}$$

7.3 Controlo Adaptativo multivariado

Na maioria dos sistemas de controlo por realimentação, pequenos desvios nos valores dos parâmetros projectados não causarão quaisquer problema na operação normal do sistema, uma vez que eles são linearizados para funcionar numa determinada zona que se pode considerar linear nessa vizinhança. No entanto, se os parâmetros variam muito, devido por exemplo a condições ambientais ou mesmo porque o processo não é linear, e portanto ao sair dessa zona actuação o processo deixa de ter um desempenho satisfatório

Se a função de transferência do processo puder ser identificada continuamente, então pode-se compensar as variações da função de transferência do processo variando os parâmetros ajustáveis do controlador, obtendo-se desta forma um desempenho satisfatório do sistema continuamente nas diversas condições de operação. Esta abordagem adaptativa é bastante útil para lidar com problemas onde o processo é normalmente exposto a condições de operação variáveis, de tal forma que os parâmetros do processo mudam de tempos em tempos.

Um sistema de controlo adaptativo é aquele que mede, de forma contínua e automática, as características dinâmicas, compara-as com as características dinâmicas desejadas e usa a diferença para variar parâmetros ajustáveis do sistema (normalmente características do controlador) ou para gerar um sinal actuante de tal forma que o desempenho óptimo possa ser mantido independentemente das condições ambientais; alternativamente, tal sistema pode medir continuamente o seu próprio desempenho de acordo com um dado índice de desempenho e modificar, se necessário, os seus próprios parâmetros, de tal forma a manter desempenho óptimo independentemente das mudanças ambientais (Ogata K. , 1970).

A base do controlo adaptativo está na premissa de que há alguma condição de operação ou desempenho para o sistema melhor do que qualquer outra. Portanto torna-se necessário definir o que constitui um desempenho óptimo. Em sistemas de controlo adaptativo, tal desempenho é definido em termos do índice de desempenho que normalmente está indexado à minimização do custo de operação, ou à maximização do lucro.

O índice de desempenho deve ser confiável, ou deve ser uma medida uniforme de “qualidade” para sistemas de todas as ordens. Deve ter selectividade, ou envolver um óptimo definido com precisão em função dos parâmetros do sistema. Não devem ter óptimos locais e deve ser facilmente aplicável para sistemas práticos e facilmente mensurável.

De notar que em geral todos os índices de desempenho matematicamente tratáveis, inclusive os quadráticos, têm a desvantagem (séria) de não darem informação sobre as características de resposta transitória do sistema, pelo que, um sistema que é projectado para operar de forma óptima do ponto de vista de máximo “lucro” pode ter características transitórias indesejáveis, ou mesmo ser instável. Desta forma, para assegurar características de resposta transitória satisfatórias, pode-se necessitar de critérios secundários relacionados com a características de resposta para poder influenciar na escolha dos elementos de ponderação do custo.

O índice de desempenho usado num sistema de controlo adaptativo define o desempenho óptimo para aquele sistema. Isto significa que o índice de desempenho essencialmente dá o limite superior do desempenho do sistema, pelo que a selecção de um índice de desempenho adequado é fundamental.

Um controlador adaptativo pode consistir nas seguintes funções:

1. Identificação das características dinâmicas do processo
2. Decisão baseada na identificação do processo
3. Modificação ou actuação baseada na decisão tomada

7.3.1 Modelo base de estratégia de controlo adaptativo multi-input multi-output

Grande parte dos processos modernos de produção é controlada por controladores automáticos, normalmente tipo PID. Os controladores automáticos podem gerir sistemas bastantes complexos de entradas e saídas múltiplas, no entanto, um problema comum nestes controladores é a necessidade de ajustamento ou sintonização se as condições de operação, especificações ou algumas variáveis externas se alteram.

Este tema tem merecido as atenções dos mais diversos ramos da engenharia no sentido de reduzir a variabilidade e aumentar a eficácia dos processos e simultaneamente reduzir a dependência do sempre subjacente factor humano, quer do operador quer do gestor do processo. Para atingir estes objectivos, nos últimos anos, tem sido desenvolvidos esforços conjuntos das áreas do controlo estatístico e engenharia de controlo, surgindo assim propostas de estratégias de monitorização e controlo que simultaneamente tentam optimizar as características da qualidade inerentes a processo, compensar o efeito das perturbações internas e externas e detectar oportunidades.

Nos últimos anos, a indústria mais pujante em termos de crescimento e performance no desenvolvimento tecnológico dos meios de produção, tem sido a indústria de semicondutores. Consequentemente, as propostas mais recentes neste âmbito tem tido como objectivo este tipo de industria, mais propriamente a área de run-to-run (R2R), (del Castillo, 1996), (Del Castillo & Yeh, 1998). Todos os estudos subsequentes, (Jen, Jiang, & Fan, 2004), (Chang, Hao, & Baras, 2000) entre outros, seguiram mais ou menos o mesmo modelo e para os mesmos fins, introduzindo por vezes novos recursos tais como redes neuronais ou lógica fuzzy. Neste trabalho tenta-se partir do esquema base (integração EPC/SPC solução de controlo) apresentado (Del Castillo, 1998), e extrapolá-lo para a generalidade dos processos, sejam eles discretos ou contínuos, multivariados ou univariados.

7.3.1.1 Controlo run-to-run

Recentemente, a área de controlo *run.to-run* tem recebido considerável interesse literário, principalmente no segmento de produção de semicondutores (Del Castillo, 1998). O esquema de controlo R2R tem sido tipicamente aplicado a uma etapa na produção de semicondutores, onde as variáveis de entrada de um determinado equipamento são calculadas baseadas em medidas tiradas entre lotes "*batches*" de *wafers* de silicone. Este tipo de controlo é considerado um controlador supervisor que regula o ajustamento (*set-point*) entre lotes na tentativa de ajustar a saída do processo aos valores de referência e simultaneamente reduzir a variabilidade.

Os sistemas multi-input multi-output são os mais frequentemente encontrados em situações práticas na indústria de semicondutores. Um problema comum é que as aplicações R2R mais evoluídas de técnicas de controlo adaptativo destinam-se a controlo de modelos

lineares para processos MIMO. No entanto, têm sido feitos alguns estudos em modelo não lineares. Se a superfície da resposta real do processo for mais ou menos plana em toda a região operacional, então um controlador linear deverá dar bons resultados. No entanto, os efeitos das não linearidades das variáveis de entrada ajustáveis na relação funcional e a dinâmica de autocorrelação entre as saídas do processo ocorrem bastante frequentemente na produção de semicondutores. Para ultrapassar estas dificuldades, (del Castillo, 1996) apresentou um modelo de controlador auto-sintonizante derivado da filosofia dos controladores de variância mínima para lidar com os processos de ganho puro múltiplo acoplados com complexos modelos de distúrbio. Adicionalmente, implementou uma carta de controlo EWMA multivariada para monitorizar as saídas de controlo com a função de adicionar uma zona morta ao controlador com o intuito de conjugar a variabilidade das saídas com a redução das alterações nas variáveis de entrada. (Del Castillo & Yeh, 1998) propôs outro controlador MIMO R2R denominado de OAQC (*Optimizing Adaptative Quality Controller*) que era particularmente projectado para lidar com funções de transferência não lineares.

7.3.1.2 Algoritmo base do controlador OAQC

A maior parte dos controladores R2R são controladores baseados no modelo. Na prática, os modelos têm que ser estimados e inseridos no controlador, na fase de pré-produção, onde os processos são optimizados e qualificados. A qualificação determina os valores iniciais das variáveis manipuladas (varáveis de entrada). Se os valores de referência são desconhecidos, então tem que ser determinados também na fase de optimização. Nesta fase recorre-se normalmente ao desenho de experiências e à metodologia de resposta em superfície.

A optimização de sistemas de controlo tem uma característica distintiva básica: na ausência de perturbações ou ruído, as relações no estado estacionário entre as entradas e as saídas é uma função que tem um máximo ou um mínimo. O objectivo do controlo óptimo é encontrar esse ponto e manter o processo o mais perto possível desse ponto, na presença de distúrbios ou deriva.

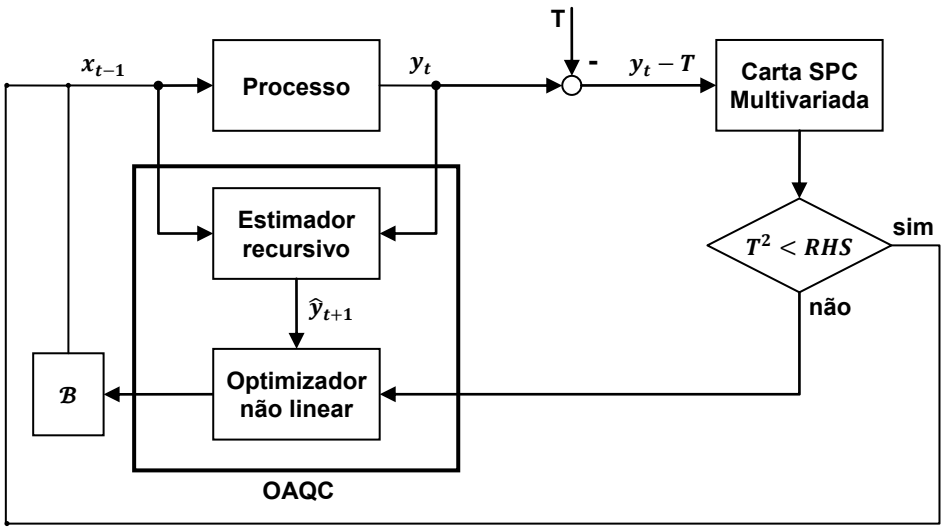


Figura 7-4 Controlador OAQC com carta SPC actuando como zona morta

O sistema de controlo OAQC assume que o comportamento do processo com $\boldsymbol{n}$ entradas e $\boldsymbol{p}$ saídas pode ser modelado de acordo com uma função de segunda ordem da forma

$$\boldsymbol{(I_p - LB)y_t = y(0) + Nz_{t-1} + Mt + (I_p - CB)\varepsilon_t} \quad (7.9)$$

Onde $z_t^{'} = \left[\boldsymbol{x_t},\ \boldsymbol{x_t^2},\ x_t^{(i)}x_t^{(j)}(i<j)\right]$ é um vector de dimensão $2n + (n\,(n-1)/2)$ que contém a expansão quadrática do vector $\boldsymbol{x_t}$ das variáveis de entrada, $\boldsymbol{y_t}$ é um vector de dimensão p das características da qualidade, $\boldsymbol{y(0)}$ é um vector de dimensão p que contém os valores de offset, $\boldsymbol{x_{t-1}}$ é um vector de dimensão n dos factores controláveis, $\boldsymbol{t}$ é o vector que contém o índice t nos seus p componentes e $\{\varepsilon_t\}$ é uma sequência de ruído branco multivariado. Este modelo é suficientemente geral para permitir a modelação da maior parte dos distúrbios e não linearidades aproximadas que parecem ocorrer nas aplicações de controlo R2R para os processos de produção de semicondutores. O modelo permite ainda alguma dinâmica.

Os parâmetros $\hat{L}$, $\hat{M}$ e $\hat{N}$ são estimados on-line recorrendo a algoritmos dos mínimos quadrados recursivos (ver secção 4.3.3).

Optimização não linear
A optimização dos valores das variáveis de entrada é efectuada com recurso ao índice de performance (função objectivo - Figura 7-5) da versão multivariada do controlador de Clarke e Gawthrop um passo à frente (secção 6.2.3) dado por

$$\boldsymbol{J = \left[(\hat{y}_{t+1|t} - T)'W\,(\hat{y}_{t+1|t} - T) + (x_t - x_{t-1})\Gamma(x_t - x_{t-1})\right]} \quad (7.10)$$

Que é minimizado sujeito às restrições na entrada e na saída:

$$\boldsymbol{L_x \le x_t \le U_x}$$
$$\boldsymbol{L_y \le y_t \le U_y}$$

em que os vectores $\boldsymbol{L_x}$, $\boldsymbol{U_x}$, $\boldsymbol{L_y}$ e $\boldsymbol{U_y}$ são os valores que limitam as zonas de operação das entradas e saídas do processo. A optimização de J em ordem aos factores controláveis $\boldsymbol{x_t}$ é realizada com recurso ao cálculo numérico através ao método "*Penalty-Barrier*". O termo "*Barrier*" é aplicado às variáveis de entrada e garante uma solução com os valores dentro dos limites.

Carta EWMA multivariada
Uma carta EWMA (carta de médias moveis exponencialmente amortecidas) (Lowry, Woodall, Champ, & Rigdon, 1992) é adicionada ao anel de controlo com a dupla função de reduzir a variabilidade das variáveis de entrada ou monitorizar as saídas do processo.

Actuando como limitadora de zona morta (Figura 7-2) implica que os parâmetros do modelo interno do controlador e as variáveis de entrada apenas sofrem alterações quando é detectada uma situação em que as saídas do processo se afastam significativamente dos valores de referência, ou seja, caem fora dos limites da carta EWMA.

Actuando como monitor do vector média do processo, detecta situações de alterações da média do processo que controlador não consegue compensar. Estas situações são normalmente provocadas pela saturação dos actuadores (variáveis de entrada).

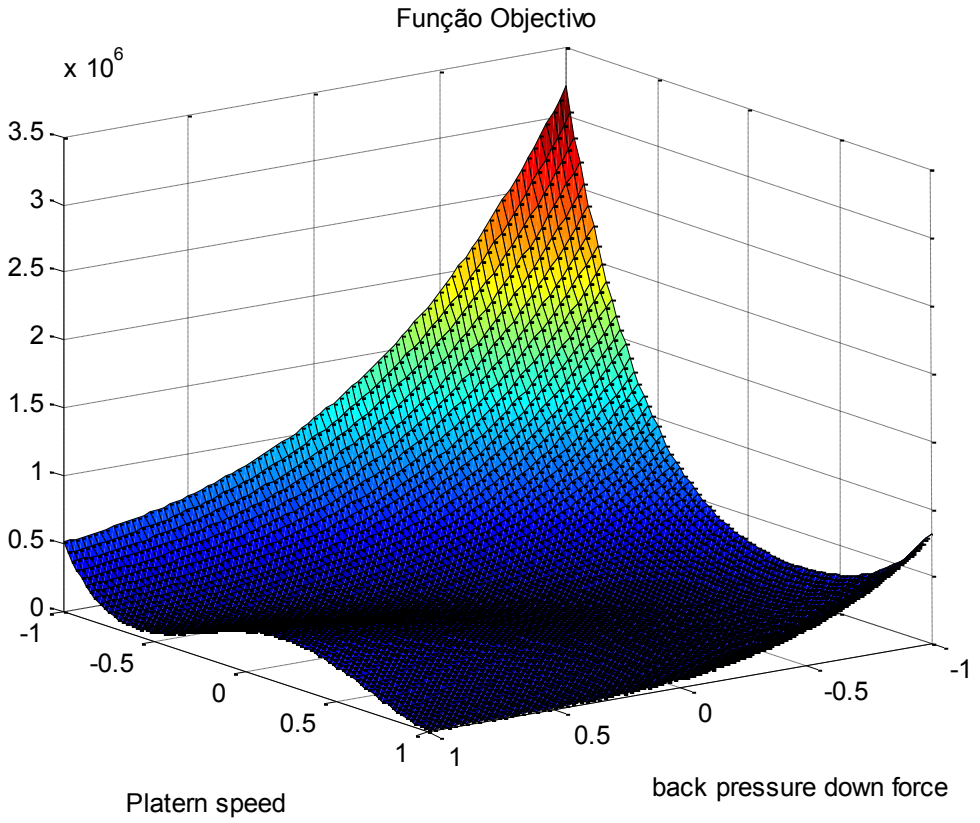


Figura 7-5 Exemplo de função de custo de controlador R2R com varáveis de entrada normalizadas

8 Desenvolvimentos Práticos

Ao contrário de o que estava previsto na data da decisão da abordagem deste tema, não foi possível fazer o trabalho de campo inerente a este tema. Esse facto ficou a dever-se basicamente ao prazo apertado para a elaboração deste trabalho, e à dificuldade actual de identificar situações reais com interesse de parte a parte na abordagem deste tema. Pelo que a solução encontrada passou por executar este trabalho tendo por base dados reais, já existentes, referentes a um processo real de fluxo contínuo, disponíveis num trabalho realizado anteriormente, dentro do mesmo âmbito. Esse trabalho visou a integração da metodologia do controlo estatístico do processo com engenharia de controlo no contexto de sistemas de produção de fluxo contínuo.

Apesar da grande semelhança entre os dois temas, as abordagens são bastantes diferentes, com objectivos bastantes diferentes. Devido ao excelente trabalho feito no âmbito do tratamento dos dados em bruto, esse tema não foi abordado neste trabalho, nem tão pouco se mexeu nos dados herdados à excepção de preencher alguns valores em falta na série de dados referente à saída do processo. Pelo exposto, e devido à falta de contacto com o processo real, este processo foi tratado exactamente como se de uma caixa preta se tratasse, apesar de por vezes os dados indiciarem algumas características próprias que não foi possível confirmar.

8.1 Descrição e caracterização do processo

8.1.1 Processo base

O processo base era constituído por sistema multi-input multi-output com nove variáveis de entrada e três variáveis de saída. As variáveis exógenas (de entrada eram constituídas por seis variáveis controladas e três variáveis referentes às características da matéria prima que, no âmbito deste trabalho, se consideravam não controláveis (manipuláveis), mas cujos níveis eram conhecidos, ou seja, de acordo com o ponto 6.2.4, ter-se-ia três variáveis do tipo X_1 seis variáveis do tipo X_2 ou X_3 e três variáveis de saída (Y).

Variáveis de entrada

As variáveis de entrada são constituídas por dois tipos com características diferentes: as variáveis referentes a características intrínsecas da matéria-prima, e as variáveis de controlo. Os níveis das variáveis referentes às características intrínsecas supõem-se conhecidos à priori, por medição directa em cada instante, ou por amostragem, podendo para tal utilizar, por exemplo, cartas de controlo. As variáveis de controlo são variáveis manipuladas que se vão ajustar em cada instante de forma a compensar a dinâmica do processo, a oscilação das variáveis referentes às características da matéria-prima e outras perturbações inerentes ao processo.

As variáveis de entrada podem então ser codificadas do seguinte modo:

- **Variáveis referentes a características intrínsecas da matéria-prima:**

  - **Hm**
  - **Dn**
  - **Gr**
- **Variáveis de controlo:**
  - **AA**
  - **IS**
  - **LBC4**
  - **LBC8**
  - **TemC4**
  - **TemC8**

Variáveis de saída

As variáveis de saída referem-se a três das características da qualidade cujos níveis devem ser assegurados:

- **Visc**
- **Ref**
- **IM**

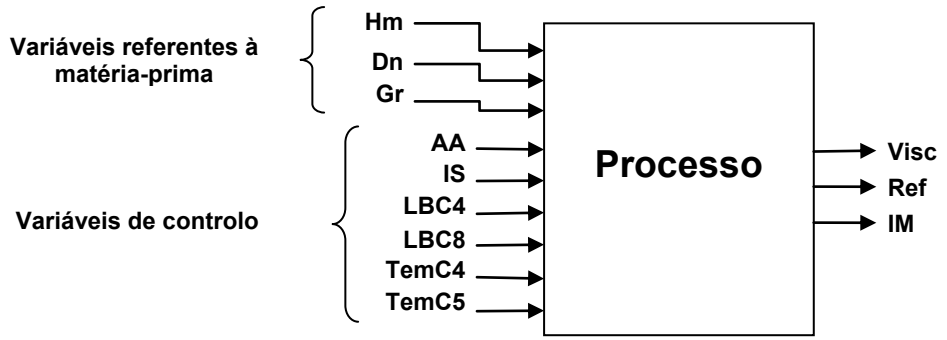


Figura 8-1 Processo base

Este seria o modelo ideal de processo a implementar, porque para além de se tratar de um sistema MIMO, as variáveis de entrada poderiam abranger os tipos existentes no modelo básico teórico (secção 6.2.4). Além disso, a integração do controlo estatístico poderia ser bastante útil para monitorizar as variáveis referentes à matéria-prima (tipo X1), e assim evitar uma medição contínua dessas variáveis. Contudo este modelo não foi possível implementar devido à falta de dados referentes às características das varáveis da matéria-prima e às variáveis de saída Ref e IM. Consequentemente, essas variáveis foram retiradas do modelo base, resultando assim o denominado modelo de estudo composto pelas seis

variáveis de controlo e uma variável de saída, resultando assim um modelo MISO (*multi-input/single-output*).

8.1.2 Processo de estudo

O processo de estudo passou então a ser caracterizado por um modelo MISO com seis variáveis de entrada:

- **AA**
- **IS**
- **LBC4**
- **LBC8**
- **TemC4**
- **TemC5**

E uma variável de saída

- **Visc**

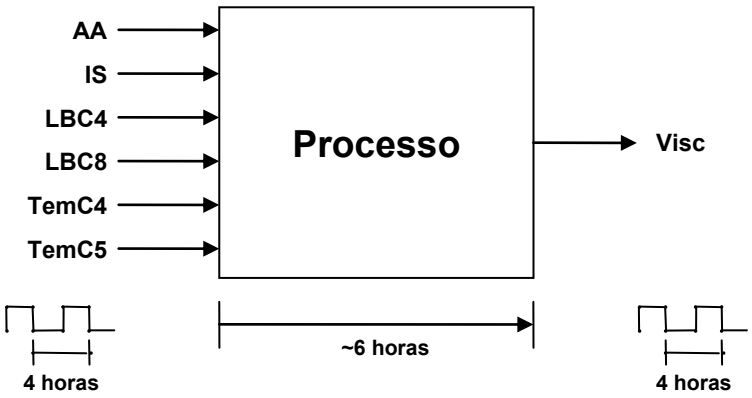


Figura 8-2 Modelo do processo de estudo

Os dados foram recolhidos com um período de amostragem de quatro horas, e o tempo médio de processamento é de cinco horas.

As séries são constituídas por 248 observações igualmente espaçadas (ANEXO III) Dos dados disponíveis verificou-se a falta de alguns valores na série da variável de saída, nomeadamente as observações número 35, 63, 83 e 134. Essas lacunas formam preenchidas com o valor correspondente à média entre a observação imediatamente anterior e a observação imediatamente posterior. Todo o trabalho de desenvolvimento prático assenta nesta série multivariada. Uma primeira análise grosseira do comportamento do processo pode ser obtida através da visualização dos gráficos das séries envolvidas (ANEXO III). Pode então verificar-se que se está perante séries estacionárias, pelo que à primeira vista será de excluir qualquer componente integrativa no modelo do processo. Para se ter uma ideia mais quantitativa das características do processo, determinaram-se

algumas estatísticas preliminares dos dados disponíveis, que são apresentadas na Tabela 8.1.

Tabela 8.1 - Estatísticas preliminares

	AA	IS	LBC4	LBC8	TemC4	TemC5	Visc
Média	0.00	0.00	0.00	0.00	0.22	0.17	1074.4
Mediana	-0.21	-0.15	-0.38	0.04	0.00	0.00	1078.0
Desvio Padrão	3.582	2.021	2.593	0.721	1.578	1.994	48.3
Máximo	12.49	6.70	6.67	3.14	4.00	7.00	1200.0
Mínimo	-12.62	-5.10	-5.33	-1.16	-4.00	-5.00	937.0
Amplitude	25.11	11.80	12.00	4.30	8.00	12.00	263.0

De notar que, por questões de sigilo relativamente ao processo em causa, os valores referentes às variáveis de entrada já se encontram centrados em zero, ou seja, os valores apresentados são as diferenças para a média.

Antes de se prosseguir com a análise de autocorrelação verificou-se a normalidade dos dados no sentido de averiguar a estabilidade do processo e a possibilidade de implementação de controlo estatístico sem ajustamento. A ferramenta utilizada nesta análise foi o teste de Kolmogorov-Smirnov.

Tabela 8.2 - Teste Kolmogorov-Smirnov

	AA	IS	LBC4	LBC8	TemC4	TemC5	Visc
P value	0,687	0,359	0,007	0,013	0,001	0,002	0,346
D	0,045	0,058	0,106	0,100	0,128	0,116	0,059
CV(5%)	0,086	0,086	0,086	0,086	0,086	0,086	0,086
H	0	0	1	1	1	1	0

A primeira linha dá o valor de percentagem correspondente ao valor da estatística D (segunda linha) para cada uma das séries. A terceira linha apresenta os valores críticos para uma significância de 5%. A quarta linha indica a hipótese que não se pode rejeitar, se a hipótese nula (valor 0) se a hipótese alternativa (valor 1). Da análise da Tabela 8.2 pode então concluir-se que as séries de entrada AA e IS não dão indícios de não normalidade de dados, ao passo que as séries de entrada LBC4, LBC8 TemC4 e TemC5 indiciam claramente um comportamento não normal. Quanto à serie de saída, para se ter uma ideia de aderência dos dados a uma distribuição normal, construiu-se um gráfico da distribuição empírica cumulativa (Figura 8-3). Numa análise preliminar aos dados de saída pode ficar-se com a ideia de que se trata de um processo controlado. Para se ter uma ideia preliminar mais precisa sobre a necessidade de implementação de uma estratégia de controlo mais elaborada, no sentido de se tentar ajustar a média do processo e reduzir a sua variabilidade, era conveniente conhecer-se as especificações do processo.

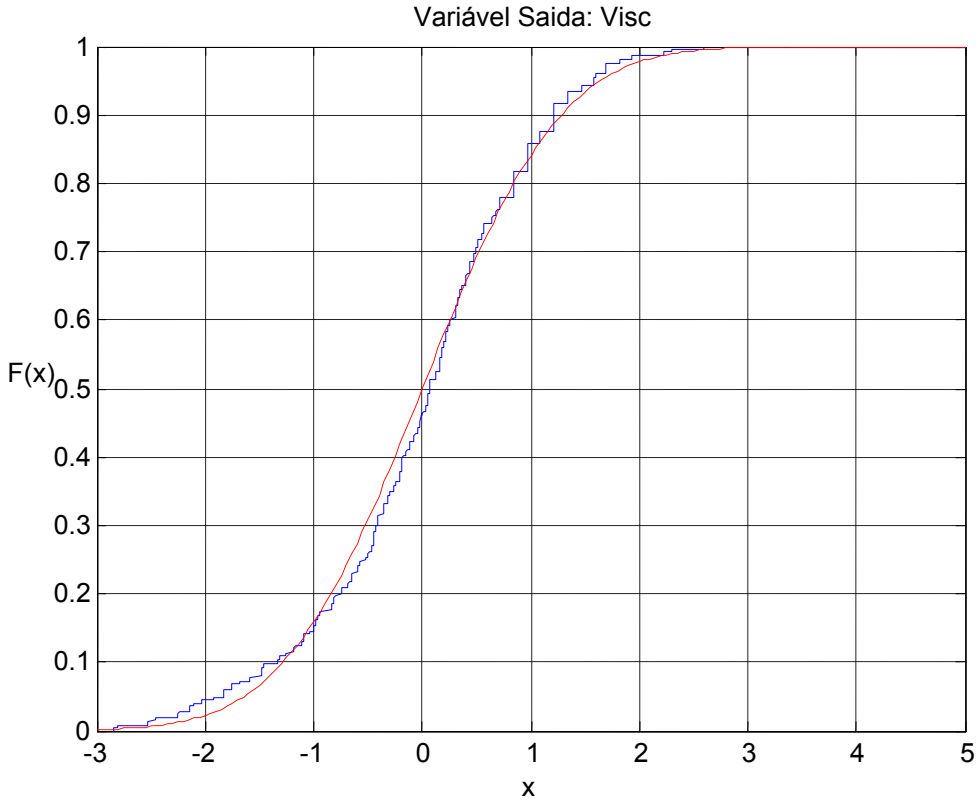


Figura 8-3 Gráfico da distribuição empírica cumulativa da variável de saída

8.1.3 Análise de correlação e coeficientes de Yule-Walker

Como a análise preliminar dos dados não indicia a existência da componente integrativa, pode-se, numa primeira fase, partir do pressuposto de que a série multivariada composta pelas sete variáveis pode ser representava por um modelo ARMA multivariado. Para se ter uma ideia muito preliminar da ordem do modelo precedeu-se à análise das funções de autocorrelação e aos coeficientes de Yule-Walker de acordo com o ponto 6.1.3.2.

Da análise das funções de autocorrecção, autocorrelação cruzada e coeficientes de Yule-Walker, fica-se com a ideia de que todas as variáveis apresentam alguma dinâmica própria, para além de indiciarem alguma autocorrelação entre elas. Outra hipótese que se levanta é a possibilidade de existência de algum controlo por realimentação, uma vez que existem algumas variáveis de entrada que denotam alguma autocorrelação com a variável de saída. O problema está directamente indexado à aquisição de dados para identificação do processos, uma vez que a própria industria, ou os “donos” dos processos, evitam “sacudir” os processos. Ou seja, para os dados de identificação reflectirem o comportamento do processo as entradas devem ser aleatórias, independentes e oscilar o suficiente para que essas alterações nas entradas tenham reflexo nas variáveis de saída. Este problema tem sido constantemente denunciado nas mais diversas bibliografias, (Box, Jenkins, & Reinsel, 2008), (Del Castillo, 2002) entre outros, pelo que algum trabalho tem sido feito ultimamente no sentido de melhorar as metodologias de identificação em malha fechada como é o caso de (Vanli & Del Castillo, 2007).

Outro facto que convém ser observado, é que nem todas as variáveis de entrada parecem ter influência na saída, ou seja parece existirem variáveis de entrada com as quais a variável de saída não está correlacionada. Isso não quer dizer que a variável de saída não depende dessa variável, poderá querer dizer é que a variável de saída não é sensível à variação verificada nessa variável de entrada. Este facto poderá evidenciar uma boa afinação do processo.

Como se pode verificar, não é fácil tirar uma conclusão, embora preliminar, sobre as ordens do modelo, embora evidencie a existência da componente autoregressiva. Verifica-se assim que uma técnica muito útil para a análise univariada torna-se pouco eficaz na análise multivariada, pelo menos neste caso.

Numa análise puramente exploratória ajustou-se um modelo autoregressivo puro de 3ª ordem sem qualquer restrição através do modelo de regressão linear como descrito no ponto 6.1.3.3 que se apresenta abaixo. Os valores entre parênteses correspondem aos respectivos desvios padrão.

$$\widehat{\boldsymbol{\Phi}}_{\mathbf{1}} = \begin{bmatrix} \mathbf{0.640} & \mathbf{-0.283} & \mathbf{0.020} & \mathbf{-0.284} & \mathbf{-0.027} & \mathbf{-0.153} & \mathbf{-0.004} \\ \mathbf{(0.073)} & \mathbf{(0.124)} & \mathbf{(0.165)} & \mathbf{(0.570)} & \mathbf{(0.187)} & \mathbf{(0.106)} & \mathbf{(0.005)} \\ \mathbf{-0.020} & \mathbf{0.523} & \mathbf{-0.119} & \mathbf{0.226} & \mathbf{0.080} & \mathbf{0.063} & \mathbf{0.000} \\ \mathbf{(0.039)} & \mathbf{(0.065)} & \mathbf{(0.087)} & \mathbf{(0.300)} & \mathbf{(0.098)} & \mathbf{(0.056)} & \mathbf{(0.002)} \\ \mathbf{-0.148} & \mathbf{0.082} & \mathbf{0.541} & \mathbf{0.415} & \mathbf{-0.238} & \mathbf{0.089} & \mathbf{0.002} \\ \mathbf{(0.034)} & \mathbf{(0.057)} & \mathbf{(0.076)} & \mathbf{(0.262)} & \mathbf{(0.086)} & \mathbf{(0.049)} & \mathbf{(0.002)} \\ \mathbf{0.004} & \mathbf{-0.015} & \mathbf{-0.030} & \mathbf{0.584} & \mathbf{0.019} & \mathbf{-0.003} & \mathbf{-0.001} \\ \mathbf{(0.008)} & \mathbf{(0.014)} & \mathbf{(0.019)} & \mathbf{(0.065)} & \mathbf{(0.021)} & \mathbf{(0.012)} & \mathbf{(0.001)} \\ \mathbf{-0.049} & \mathbf{0.008} & \mathbf{-0.030} & \mathbf{-0.193} & \mathbf{0.566} & \mathbf{0.000} & \mathbf{0.001} \\ \mathbf{(0.026)} & \mathbf{(0.044)} & \mathbf{(0.059)} & \mathbf{(0.204)} & \mathbf{(0.067)} & \mathbf{(0.038)} & \mathbf{(0.002)} \\ \mathbf{-0.078} & \mathbf{0.221} & \mathbf{0.077} & \mathbf{-0.247} & \mathbf{0.330} & \mathbf{0.316} & \mathbf{0.002} \\ \mathbf{(0.047)} & \mathbf{(0.079)} & \mathbf{(0.105)} & \mathbf{(0.363)} & \mathbf{(0.119)} & \mathbf{(0.067)} & \mathbf{(0.003)} \\ \mathbf{-1.097} & \mathbf{5.791} & \mathbf{-3.217} & \mathbf{2.875} & \mathbf{-5.148} & \mathbf{-2.639} & \mathbf{0.457} \\ \mathbf{(0.994)} & \mathbf{(1.675)} & \mathbf{(2.230)} & \mathbf{(7.718)} & \mathbf{(2.527)} & \mathbf{(1.431)} & \mathbf{(0.064)} \end{bmatrix}$$

$$\widehat{\boldsymbol{\Phi}}_{\mathbf{2}} = \begin{bmatrix} \mathbf{0.034} & \mathbf{0.045} & \mathbf{-0.022} & \mathbf{-0.442} & \mathbf{0.209} & \mathbf{0.057} & \mathbf{0.001} \\ \mathbf{(0.083)} & \mathbf{(0.146)} & \mathbf{(0.176)} & \mathbf{(0.675)} & \mathbf{(0.214)} & \mathbf{(0.112)} & \mathbf{(0.005)} \\ \mathbf{0.009} & \mathbf{0.060} & \mathbf{0.015} & \mathbf{-0.371} & \mathbf{-0.080} & \mathbf{-0.055} & \mathbf{0.001} \\ \mathbf{(0.044)} & \mathbf{(0.077)} & \mathbf{(0.092)} & \mathbf{(0.355)} & \mathbf{(0.112)} & \mathbf{(0.059)} & \mathbf{(0.003)} \\ \mathbf{0.104} & \mathbf{-0.104} & \mathbf{0.207} & \mathbf{-0.003} & \mathbf{0.247} & \mathbf{-0.033} & \mathbf{-0.001} \\ \mathbf{(0.038)} & \mathbf{(0.067)} & \mathbf{(0.081)} & \mathbf{(0.311)} & \mathbf{(0.098)} & \mathbf{(0.051)} & \mathbf{(0.002)} \\ \mathbf{-0.009} & \mathbf{-0.020} & \mathbf{-0.013} & \mathbf{0.031} & \mathbf{-0.033} & \mathbf{-0.010} & \mathbf{0.000} \\ \mathbf{(0.009)} & \mathbf{(0.017)} & \mathbf{(0.020)} & \mathbf{(0.077)} & \mathbf{(0.024)} & \mathbf{(0.013)} & \mathbf{(0.001)} \\ \mathbf{0.069} & \mathbf{-0.052} & \mathbf{0.034} & \mathbf{-0.239} & \mathbf{0.023} & \mathbf{-0.003} & \mathbf{-0.001} \\ \mathbf{(0.030)} & \mathbf{(0.052)} & \mathbf{(0.063)} & \mathbf{(0.242)} & \mathbf{(0.077)} & \mathbf{(0.040)} & \mathbf{(0.002)} \\ \mathbf{0.063} & \mathbf{-0.136} & \mathbf{-0.008} & \mathbf{-0.038} & \mathbf{-0.237} & \mathbf{0.004} & \mathbf{-0.003} \\ \mathbf{(0.053)} & \mathbf{(0.093)} & \mathbf{(0.112)} & \mathbf{(0.430)} & \mathbf{(0.136)} & \mathbf{(0.071)} & \mathbf{(0.003)} \\ \mathbf{1.753} & \mathbf{0.172} & \mathbf{-2.079} & \mathbf{-0.711} & \mathbf{1.658} & \mathbf{2.626} & \mathbf{0.058} \\ \mathbf{(1.128)} & \mathbf{(1.977)} & \mathbf{(2.378)} & \mathbf{(9.148)} & \mathbf{(2.894)} & \mathbf{(1.511)} & \mathbf{(0.069)} \end{bmatrix}$$

$$\widehat{\boldsymbol{\Phi}}_3 = \begin{bmatrix} -0.092 & 0.324 & -0.049 & 0.528 & -0.220 & -0.164 & 0.001 \\ (0.075) & (0.131) & (0.151) & (0.550) & (0.198) & (0.105) & (0.004) \\ -0.005 & 0.151 & 0.090 & 0.093 & 0.075 & -0.064 & 0.003 \\ (0.040) & (0.069) & (0.079) & (0.289) & (0.104) & (0.055) & (0.002) \\ 0.060 & 0.011 & 0.099 & -0.136 & -0.155 & 0.059 & -0.001 \\ (0.035) & (0.060) & (0.070) & (0.253) & (0.091) & (0.048) & (0.002) \\ -0.008 & 0.010 & 0.008 & 0.199 & -0.053 & 0.029 & -0.001 \\ (0.009) & (0.015) & (0.017) & (0.063) & (0.023) & (0.012) & (0.001) \\ -0.071 & -0.047 & -0.077 & 0.129 & 0.098 & -0.033 & -0.002 \\ (0.027) & (0.047) & (0.054) & (0.197) & (0.071) & (0.038) & (0.002) \\ -0.069 & 0.020 & -0.086 & 0.188 & -0.143 & 0.151 & -0.004 \\ (0.048) & (0.083) & (0.096) & (0.350) & (0.126) & (0.067) & (0.003) \\ -0.202 & -1.488 & 3.504 & 5.184 & 3.890 & 0.315 & -0.036 \\ (1.020) & (1.773) & (2.047) & (7.455) & (2.688) & (1.419) & (0.060) \end{bmatrix}$$

$$\widetilde{\boldsymbol{\Sigma}} = \begin{bmatrix} 7.156 & 0.805 & -1.737 & 0.201 & -0.018 & 0.005 & -0.258 \\ 0.805 & 1.981 & -0.347 & 0.033 & 0.100 & 0.130 & -0.404 \\ -1.737 & -0.347 & 1.518 & -0.130 & -0.063 & 0.038 & -4.800 \\ 0.201 & 0.033 & -0.130 & 0.093 & 0.012 & 0.014 & -0.438 \\ -0.018 & 0.100 & -0.063 & 0.012 & 0.920 & 0.575 & -1.027 \\ 0.005 & 0.130 & 0.038 & 0.014 & 0.575 & 2.900 & -3.483 \\ -0.258 & -0.404 & -4.800 & -0.438 & -1.027 & -3.483 & 1313.900 \end{bmatrix}$$

A matriz $\widetilde{\boldsymbol{\Sigma}}$ é a matriz de covariância dos resíduos após o ajustamento por este modelo VAR(3). Uma das características que as matrizes $\boldsymbol{\Phi's}$ evidenciam é que as seis primeiras linhas da sétima coluna não são significativas ao nível de dois desvios padrão, pelo que estes valores deverão ser nulos. Esse facto sugere que, ao contrário da hipótese levantada anteriormente, que as variáveis de entrada não são influenciadas pela variável de saída, o que indicia a não existência de controlo por realimentação. Outro conclusão que se pode tirar da observação das matrizes $\widehat{\boldsymbol{\Phi}}_2$ e $\widehat{\boldsymbol{\Phi}}_3$, é que existem poucos coeficientes significativos o que indica que os componentes destas matrizes, caso se conclua que a ordem do modelo é pelo menos 3, serão maioritariamente nulos.

8.2 Identificação do processo

8.2.1 Determinação das ordens do modelo

Para determinar as ordens *p* e *q* do modelo ARMA seguiu-se a metodologia descrita no ponto 6.1.3.

Inicialmente procedeu-se ao ajuste a um modelo autoregressivo puro para as ordens $m = 1, \dots, 8$ pelo método dos mínimos quadrados, determinando-se para cada ordem o valor M_n da estatística de teste LR e respectiva percentagem, os valores dos critérios de selecção de modelo AIC e HQ, bem como o determinante da matriz de covariância dos respectivos resíduos (Tabela 8.3).

Tabela 8.3 - Sumário de resultados de ajustamento a modelo autoregressivo para diversa ordens

m (Ordem AR)	1	2	3	4	5	6	7	8
$\lvert\tilde{\Sigma}\rvert$	7680.8	5120.8	3527.5	2814.7	1986.6	1481.9	1064.2	703.9
AIC_m	9.343	**9.338**	9.368	9.549	9.611	9.731	9.816	9.823
HQ_m	**9.624**	9.900	10.214	10.681	11.029	11.438	11.815	12.114
Estatistica M_m	1220.2	97.611	79.865	44.228	72.380	55.365	62.278	66.814
P-value	**1.0000**	**1.0000**	**0.9965**	0.3334	0.9834	0.7529	0.9036	0.9539

Estes resultados indicam que para um modelo AR puro, um modelo de segunda ordem será talvez o mais indicado, embora não se excluam as hipóteses de primeira e terceira ordem.

Considerou-se também a possibilidade de um modelo misto ARMA. Recorrendo a métodos de estimativa dos mínimos quadrados, num primeiro estágio determinaram-se os resíduos de um modelo AR de ordem 8 ($m^* = 8$) e respectiva matriz de covariância. Num segundo estágio ajustaram-se vários modelo ARMA(*p,q*) para diversas combinações de *p* e *q*, recorrendo a estimativas dos mínimos quadrados, determinando-se para cada combinação os critérios AIC (Tabela 8.4) e AQ (Tabela 8.5).

Tabela 8.4 Critério AIC para selecção da ordem do modelo ARMA

	q					
p	**0**	**1**	**2**	**3**	**4**	**5**
0	13.613	13.236	13.315	13.519	13.753	14.105
1	9.407	9.558	9.891	10.218	10.527	10.954
2	9.587	9.788	10.112	10.480	10.866	11.328
3	9.853	10.088	10.433	10.826	11.268	11.693
4	10.311	10.535	10.907	11.271	11.757	12.156
5	10.599	10.941	11.311	11.681	12.155	12.515

Tabela 8.5 Critério HQ para selecção da ordem do modelo ARMA

	q					
p	**0**	**1**	**2**	**3**	**4**	**5**
0	13.613	13.523	13.889	14.384	14.909	15.554
1	9.693	10.131	10.752	11.370	11.972	12.694
2	10.161	10.649	11.261	11.921	12.600	13.357
3	10.717	11.241	11.874	12.555	13.291	14.012
4	11.467	11.980	12.641	13.294	14.069	14.765
5	12.048	12.681	13.341	14.001	14.764	15.414

O critério AIC aponta para um modelo AR de primeira ordem, não sendo de excluir a segunda ou terceira ordem, ou mesmo um modelo misto até três parâmetros. O critério HQ aponta mais ou menos para os mesmos resultados, embora exista uma diferença mais acentuada entre modelos de dois parâmetros para três parâmetros. Se a decisão fosse tomada apenas com base nestes critérios, a decisão cairia num modelo AR(1).

Para se ter uma ideia mais precisa sobre a ordem da componente autoregressiva a seleccionar optou-se por determinar os três modelos ($p = 1, 2, 3$) com restrição de parâmetros seguindo o procedimento descrito no ponto 6.1.3.5. Para cada modelo determinou-se os indicadores AIC e HC, que privilegiam o menor numero de parâmetros, bem como a estatísticas Q_s (expressões (6.63) e (6.64)). Para além destes indicadores, determinou-se ainda o valor de R^2 dado por

$$\boldsymbol{R_y^2 = \frac{\sum_{t=1}^{N} \hat{y}_t^2}{\sum_{t=1}^{N} y_t^2} = 1 - \frac{\sum_{t=1}^{N} \hat{\varepsilon}_t^2}{\sum_{t=1}^{N} y_t^2}} \quad (8.1)$$

que permite medir a percentagem de variação (da variável de saída) explicada pelo modelo.

Tabela 8.6 Sumario dos resultados obtidos no ajustamento a modelos AR para ordens 1, 2 e 3

Ordem	*1*	*2*	*3*
AIC	9,4981	9,5713	9,8228
HQ	9,7784	10,1340	10,6690
$\lvert\tilde{\Sigma}\rvert$	8967,3	6467,8	5556,8
$\boldsymbol{Q_s}$	180,28	198,2	195,03
P-value $(\boldsymbol{Q_s})$	1	1	1
$\boldsymbol{R^2}$	0,36638	0,40763	0,41855

Os resultados obtidos estão sumarizados na Tabela 8.6. como se pode verificar, os indicadores AIC, HQ e Q_s privilegiam o modelo de primeira ordem, ao passo que os indicadores $\lvert\tilde{\Sigma}\rvert$ e R^2 apontam para o modelo de terceira ordem. Como o indicador Q_s não exclui o modelo de terceira ordem e este demonstra uma melhor adaptação aos dados existentes, optou-se por seleccionar esta ordem caracterizado pelos seguintes parâmetros

$$\boldsymbol{\Phi_1} = \begin{bmatrix} 0,623 & -0,207 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0,599 & 0 & 0 & 0 & 0 & 0 \\ -0,145 & 0 & 0,509 & 0,329 & -0,205 & 0 & 0 \\ 0 & 0 & -0,007 & 0,677 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0,630 & 0 & 0 \\ -0,065 & 0,140 & 0 & 0 & 0,325 & 0,329 & 0 \\ 0 & 4,550 & -3,347 & 0 & -4,326 & -2,182 & 0,474 \end{bmatrix}$$

$$\boldsymbol{\Phi_2} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0,128 & 0 & 0,240 & 0 & 0,103 & 0 & 0 \\ 0 & -0,025 & 0 & 0 & 0 & 0 & 0 \\ 0,029 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0,309 & 0 & 0 \\ 1,401 & 0 & 0 & 0 & 0 & 2,823 & 0 \end{bmatrix}$$

$$\boldsymbol{\Phi_3} = \begin{bmatrix} 0 & 0.276 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.181 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.094 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.180 & 0 & 0.019 & 0 \\ -0.059 & 0 & -0.052 & 0 & 0.149 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.175 & 0 \\ 0 & 0 & 2.074 & 0 & 3.003 & 0 & 0 \end{bmatrix}$$

$$\tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} 7,54730 & 0,85330 & -1,84160 & 0,21587 & 0,21587 & 0,05440 & -0,60310 \\ 0,85330 & 2,10580 & -0,38448 & 0,04052 & 0,11526 & 0,08286 & -0,66118 \\ -1,84160 & -0,38448 & 1,62850 & -0,13215 & -0,06446 & 0,05567 & -4,93230 \\ 0,21587 & 0,04052 & -0,13215 & 0,10289 & 0,01712 & 0,02166 & -0,65976 \\ 0,04313 & 0,11526 & -0,06446 & 0,01712 & 1,00310 & 0,64159 & -1,14130 \\ 0,05440 & 0,08286 & 0,05567 & 0,02166 & 0,64159 & 3,03430 & -3,24570 \\ -0,60310 & -0,66118 & -4,93230 & -0,65976 & -1,14130 & -3,24570 & 1364,70 \end{bmatrix}$$

A Figura 8-4 apresenta-se um gráfico das correlações cruzadas e autocorrelação dos resíduos referentes ao modelo. O gráfico referente à função de autocorrelação e correlação cruzada apresenta os limites referentes à significância de 2σ, ou seja $\pm 2\sqrt{N}$ (os gráficos referentes às variáveis de entrada encontram-se no ANEXO IV).

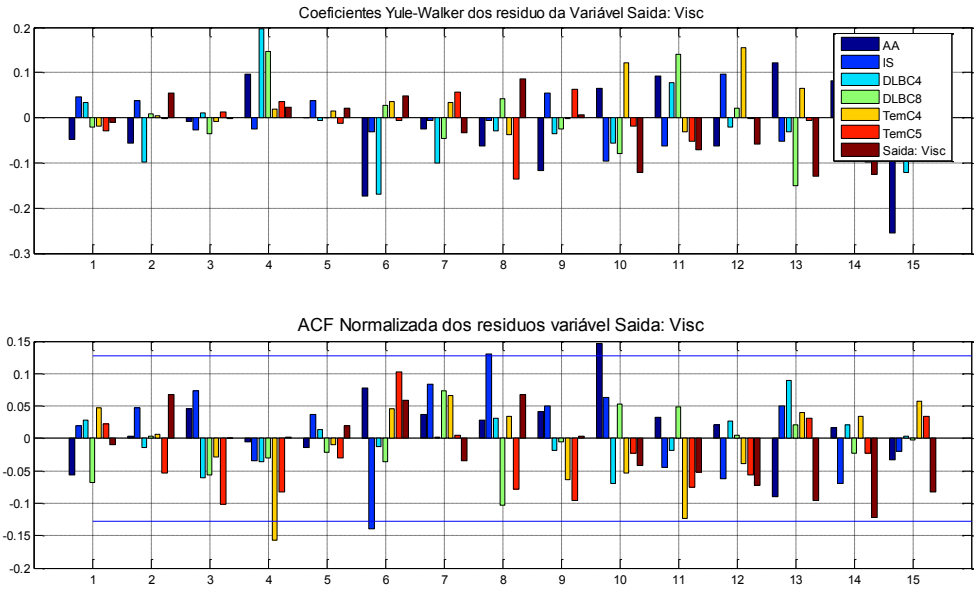


Figura 8-4 Coeficientes de Yule-Walker e ACF dos resíduos

A Figura 8-4 indicia a existência de correlação (marginal) entre a variável de saída e a variáveis *TemC4* na *lag* 4 e *IS* na *lag* 6. Esta característica poderá estar relacionada com uma fraca estrutura sazonal que ainda se mantém e que poderá ser provocada por alguma manutenção periódica ou a outro motivo idêntico. Quanto à situação verificada na *lag* 6, considera-se, no âmbito deste trabalho, que esta está suficientemente distante para se considerar significativa; de notar que se utilizasse o limite 3σ ela não seria significativa. Para colmatar a situação verificada na *lag* 4 pode-se incluir no modelo uma componente média móvel exactamente na *lag* 4 (Reinsel, 1997), passando o processo a representado por um modelo ARMA com a seguinte estrutura

$$\boldsymbol{Z_t = \Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + \Phi_3 Z_{t-3} + \varepsilon_t - \Theta_4 \varepsilon_{t-4}} \quad (8.2)$$

Procedendo à análise à matriz de covariância dos resíduos, mais precisamente à ultima linha, verifica-se que existe pelo menos uma componente γ_{73} que poderá indiciar alguma correlação cruzada instantânea. É conveniente relembrar que a periodicidade da amostragem é de 4 horas e o tempo médio de permanência dentro do sistema é de seis horas.

Isso significa não se deve colocar a hipótese de um modelo ARMA com a seguinte estrutura:

$$\boldsymbol{(I_k - \Phi_0) Z_t = \Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + \Phi_3 Z_{t-3} + \varepsilon_t - \Theta_4 \varepsilon_{t-4}} \quad (8.3)$$

Em que a matriz os elementos da matriz Φ_0 serão nulos, à excepção dos seis primeiros elementos da sétima linha. Ao modelo da expressão (8.2) atribui-se a designação sugestiva de modelo ARMA(1.2.3, 4) e, seguindo o mesmo raciocínio, ao modelo da expressão **(8.3**) atribui-se a designação de modelo ARMA(0.1.2.3, 4)

8.3 Determinação do modelo final

8.3.1 Modelo ARMA(1.2.3, 4)

Uma vez determinada a estrutura do modelo, o próximo passo é determinar os respectivos parâmetros que, como foi referido na secção 6.1.4, deverá ser através da à estimativa da máxima verosimilhança. Para se utilizar o algoritmo recursivo apresentado no ponto 6.1.4.1 (Determinação iterativa da estimativa condicional MLE), é necessário determinar um modelo inicial. Para se determinar o modelo inicial procedeu-se de acordo com o ponto 6.1.4.2. Os resultados globais do ajustamento deste modelo podem ser consultados no ANEXO V.

$$\widetilde{\boldsymbol{\Sigma}} = \begin{bmatrix} 7,0169 & 0,7630 & -0,5399 & 0,1771 & 0,0562 & -0,0372 & 12,5800 \\ 0,7630 & 2,1064 & -0,3934 & 0,0323 & 0,1105 & 0,0514 & -0,5399 \\ 0,7630 & -0,3934 & 1,5559 & -0,1251 & -0,0909 & 0,0900 & 1,8652 \\ -1,6300 & 0,0323 & -0,1251 & 0,0924 & 0,0191 & 0,0143 & 1,8652 \\ 0,0562 & 0,1105 & 0,0909 & 0,0191 & 0,0191 & 0,6228 & -0,5496 \\ -0,0372 & 0,0514 & 0,0900 & 0,0143 & 0,6228 & 2,9807 & -2,1247 \\ 12,5800 & -0,5399 & 1,8652 & -0,5496 & -2,1247 & -2,7526 & 1331,300 \end{bmatrix}$$

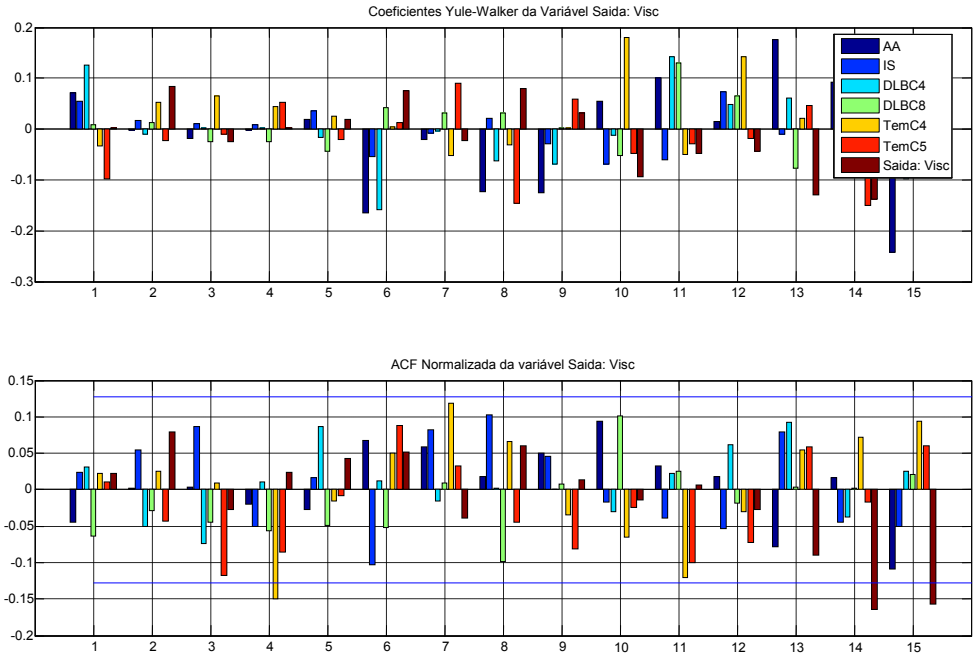


Figura 8-5 Coeficientes de Yule-Walker e ACF dos resíduos da variável de saída para o modelo ARMA(0.1.2.3, 4)

A Figura 8-5 mostra o gráfico das funções autocovariância e covariância cruzada da variável de saída. À excepção de uma correlação muito marginal na *lag* 4 com a variável *TemC4*, o modelo não indicia qualquer autocorrelação (até à *lag* 14) nem qualquer correlação cruzada (pelo menos até à *lag* 15), o que poderá à partida significar um ajustamento satisfatório. Os indicadores (Tabela 8.7) referentes ao à qualidade do ajustamento, indicam uma melhoria substancial, principalmente os dois primeiros, do modelo ARMA em relação ao modelo AR puro apesar do aumento do número de parâmetros. O indicador $\boldsymbol{Q_s}$ continua no entanto a indicar um mau ajustamento global.

O peso relativo do aumento de parâmetros em ralação à qualidade do ajustamento poderá ser maior ou menor conforme o objectivo da utilização do modelo. Como o primeiro objectivo é conseguir um modelo que represente o mais fielmente o processo, neste ponto dar-se-á maior peso à qualidade do ajustamento.

Tabela 8.7 Indicadores de ajustamento do modelo ARMA(0.1.2.3, 4)

$\lvert\widetilde{\boldsymbol{\Sigma}}\rvert$	$\boldsymbol{R^2}$	$\boldsymbol{Q_{10}}$	***P-value*** $\boldsymbol{(Q_{10})}$	***AIC***	***HQ***
4280.9	0.4328	439.88	1	9.962	11.090

Análise ao modelo ARMA obtido

Não esquecendo que se está perante um modelo ARMAX, como um dos objectivos deste trabalho passa pela identificação de modelos ARMA multivariados, optou-se por fazer também a análise do modelo obtido como se tratasse de um modelo ARMA (sem variáveis exógenas).

A Tabela 8.8 apresenta os resultados do teste de não normalidade conforme descrito no ponto 6.1.4.3.

Do ponto 6.1.4.3 e da Tabela 8.8 tem-se que $\sqrt{N/6}\,\boldsymbol{b_1} \sim \boldsymbol{N}(\mathbf{0},\mathbf{1})$, pelo que pode conclui-se que nem todas as variáveis apresentam os respectivos resíduos normalmente distribuídos. No entanto verifica-se que a variável de saída não indicia não normalidade nos seus resíduos. Do ponto 6.1.4.3 e da Tabela 8.9 tem-se que $\sqrt{N/24}\,(\boldsymbol{b_1}-\mathbf{3}) \sim \boldsymbol{N}(\mathbf{0},\mathbf{1})$, pelo que, também neste caso, o teste estatístico indicia a não normalidade global dos resíduos.

Tabela 8.8 Teste de não normalidade (*Skewness e Kurtosis*)

	AA	IS	DLBC4	DLBC8	TemC4	TemC5	Y - VISC
$\overline{\boldsymbol{v}}$	0,0018	-0,0004	-0,0007	0,0013	-0,0021	0,0000	-0,0009
$\boldsymbol{b_1}$	0,0160	0,2404	-0,2885	0,0158	-0,5286	0,0807	-0,2015
$\boldsymbol{b_2}$	3,2206	3,9301	3,6297	6,9172	4,7492	3,7759	3,2516
$\sqrt{N/6}\,\boldsymbol{b_1}$	0,1022	1,5362	-1,8435	0,1010	**-3,3776**	0,5154	-1,2877
$\sqrt{N/24}\,(\boldsymbol{b_1}-\mathbf{3})$	0,7047	2,9716	2,0118	**12,5160**	**5,5889**	2,4790	0,8040

Tabela 8.9 Teste de não normalidade global

	Valor	*P-value*
$\boldsymbol{\lambda_s}$	19,11	0,99215
$\boldsymbol{\lambda_k}$	208,05	1
$\boldsymbol{\lambda_{sk}}$	227,16	1

Apesar de este modelo satisfazer razoavelmente as pretensões, optou-se por tentar melhorá-lo, aumentando a ordem AR, no intuito de reduzir a correlação residual marginal na *lag* 4 e simultaneamente reduzir o coeficiente (7,1) da matriz de covariância residual.

8.3.2 Modelo ARMA(0.1.2.3.4, 4)

O modelo ARMA(1.2.3.4, 4) tem a mesma estrutura do modelo anterior mais uma componente autoregressiva na *lag* 4

$$(\boldsymbol{I_k}-\boldsymbol{\Phi_0})\boldsymbol{Z_t} = \boldsymbol{\Phi_1 Z_{t-1}} + \boldsymbol{\Phi_2 Z_{t-2}} + \boldsymbol{\Phi_3 Z_{t-3}} + \boldsymbol{\Phi_4 Z_{t-4}} + \boldsymbol{\varepsilon_t} - \boldsymbol{\Theta_4 \varepsilon_{t-4}} \quad (8.4)$$

O processo de ajustamento foi o mesmo que se utilizou no caso anterior, ou seja através da estimativa da máxima verosimilhança. Os resultados globais do ajustamento deste modelo podem ser consultados no ANEXO VI.

$$\widetilde{\boldsymbol{\Sigma}} = \begin{bmatrix} 6.5633 & 0.6609 & -1.5091 & 0.1387 & 0.0409 & -0.0210 & 2.6698 \\ 0.6609 & 2.0620 & -0.3214 & 0.0289 & 0.1173 & 0.0737 & -0.1525 \\ -1.5091 & -0.3214 & 1.4808 & -0.1216 & -0.0440 & 0.1039 & 3.3500 \\ 0.1387 & 0.0289 & -0.1216 & 0.0860 & 0.0164 & 0.0135 & -0.6868 \\ 0.0409 & 0.1173 & -0.0440 & 0.0164 & 1.0120 & 0.6533 & -1.2705 \\ -0.0210 & 0.0737 & 0.1039 & 0.0135 & 0.6533 & 2.9989 & -3.1212 \\ 2.6698 & -0.1525 & 3.3500 & -0.6868 & -1.2705 & -3.1212 & 1268.80 \end{bmatrix}$$

$$\widetilde{\boldsymbol{V}} = \begin{bmatrix} 2.5619 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.2580 & 1.4126 & 0 & 0 & 0 & 0 & 0 \\ -0.5891 & -0.1199 & 1.0580 & 0 & 0 & 0 & 0 \\ 0.0541 & 0.0105 & -0.0836 & 0.2756 & 0 & 0 & 0 \\ 0.0160 & 0.0801 & -0.0236 & 0.0463 & 1.0013 & 0 & 0 \\ -0.0082 & 0.0537 & 0.0997 & 0.0786 & 0.6470 & 1.6004 & 0 \\ 1.0421 & -0.2983 & 3.7126 & -1.5598 & -1.1020 & -1.6441 & 35.321 \end{bmatrix}$$

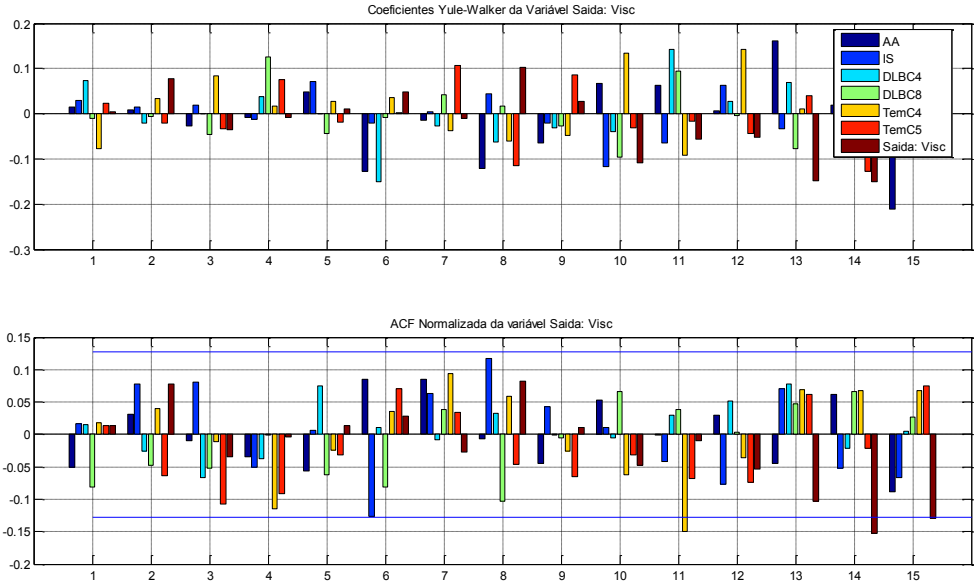


Figura 8-6 Coeficientes de Yule-Walker e ACF dos resíduos da variável de saída para o modelo ARMA(0.1.2.3.4, 4)

A Figura 8-6 mostra o gráfico das funções autocovariância e covariância cruzada da variável de saída. Como se pode verificar, o modelo não indicia qualquer autocorrelação (até à lag 14) nem qualquer correlação cruzada (pelo menos até à *lag* 15), o que poderá à partida significar um ajustamento satisfatório. No que se refere à matriz de covariância residual, verifica-se também uma ligeira melhoria.

Comparativamente, os indicadores (Tabela 8.10) referentes ao à qualidade do ajustamento, indicam alguma melhoria, principalmente os três primeiros, do último modelo em relação ao anterior.

Tabela 8.10 Indicadores de ajustamento dos modelos ARMA(0.1.2.3, 4) e ARMA(0.1.2.3.4, 4)

Modelo	$\lvert\widetilde{\Sigma}\rvert$	R^2	Q_{10}	*P-value* (Q_{10})	*AIC*	*HQ*
ARMA(0.1.2.3, 4)	4280.9	0.4328	439.88	1	9.962	11.090
ARMA(0.1.2.3.4, 4)	3567,1	0,4464	417,41	1	10,188	11,602

De notar que os elementos da última linha do parâmetro $\boldsymbol{\Phi_4}$ são todos nulos, ou seja, a ordem do modelo em relação à variável de saída não aumentou. Isso significa que o melhor ajustamento que se verifica em termos da resposta do processo se deve fundamentalmente à melhor identificação das relações dinâmicas entre as restantes variáveis do processo.

Análise ao modelo ARMA obtido

A exemplo do que se fez para o modelo ARMA(0.1.2.3, 4), também para este modelo se procedeu à análise do modelo obtido como se tratasse de um modelo ARMA (sem variáveis exógenas).

A Tabela 8.11 apresenta os resultados do teste de não normalidade conforme descrito no ponto 6.1.4.3.

Tabela 8.11 Teste de não normalidade (Skewness e Kurtosis)

	AA	IS	DLBC4	DLBC8	TemC4	TemC5	Y - VISC
$\overline{\boldsymbol{v}}$	-0.0013	-0.0013	0.0004	0.0038	0.0000	-0.0010	-0.0009
$\boldsymbol{b_1}$	0.1469	0.1741	-0.2352	-0.0630	-0.5527	0.0878	-0.3376
$\boldsymbol{b_2}$	3.2490	3.9190	3.7200	6.3583	4.6846	3.7845	3.0659
$\sqrt{N/6}\,\boldsymbol{b_1}$	0.9368	1.1102	-1.4999	-0.4015	**-3.5246**	0.5602	-2.1526
$\sqrt{N/24}\,(\boldsymbol{b_1}-\boldsymbol{3})$	0.7938	2.9301	2.2956	**10.7080**	**5.3715**	2.5013	0.2100

Tal como no modelo anterior, também para este modelo se pode concluir que nem todas as variáveis apresentam os respectivos resíduos normalmente distribuídos, nomeadamente a variável DLBC8 e TemC4. A variável de saída, neste caso, também não indicia não normalidade nos seus resíduos. Em relação ao teste global, os resultados também não se alteram significativamente em relação ao modelo anterior (Tabela 8.12).

Tabela 8.12 Teste de não normalidade global

	Valor	*P-value*
$\boldsymbol{\lambda_s}$	19,11	0,99215
$\boldsymbol{\lambda_k}$	208,05	1
$\boldsymbol{\lambda_{sk}}$	227,16	1

8.3.2.1 Conclusões

Em relação ao modelo obtido, foi o modelo possível com os dados disponíveis, mas não se pode concluir que seja bom. O valor de $\boldsymbol{Q_s}$ indica um mau ajustamento e o valor de $\boldsymbol{R^2}$ indica que o modelo apenas explica 44,6% da variabilidade. Contudo tem que se ter em conta que o modelo é construído em cima dos dados disponíveis e que estes nem sempre são gerados por processos ARMA ou ARIMA.

Uma das hipóteses aventadas no inicio do estudo deste processo era que o modelo poderia ser simplesmente descrito por um modelo univariado, controlando apenas a variável de saída (Viscosidade) com recurso ao controlo estatístico. Ou seja o modelo poderia ter a forma, denominada aqui como "*modelo zero*":

$$\boldsymbol{y_t = \mu + \varepsilon_t}$$

Procedeu-se ao teste de Kolmogorov-Smirnov e conclui-se que não se podia afirmar que os dados de saída eram não normais. Chegados a este ponto, pensou-se que era de todo pertinente comparar os dois modelos.

Tal como se procedeu no inicio em relação ao modelo zero, verificou-se a normalidade residual do modelo ARMA(0.1.2.3.4, 4) com recurso ao teste de Kolmogorov-Smirnov.

Tabela 8.13 Teste Kolmogorov-Smirnov aplicado aos resíduos do modelo ARMA(0.1.2.3.4, 4)

	AA	IS	LBC4	LBC8	TemC4	TemC5	Visc
P- value	0,502	0,480	0,461	0,103	0,206	0,678	0,835
D	0,052	0,053	0,054	0,077	0,068	0,046	0,039
CV(5%)	0,086	0,086	0,086	0,086	0,086	0,086	0,086
H	0	0	0	0	0	0	0

A primeira linha dá o valor de percentagem correspondente ao valor da estatística D (segunda linha) para cada uma das séries. A terceira linha apresenta os valores críticos para uma significância de 5%. A quarta linha indica a hipótese que não se pode rejeitar, se a hipótese nula (valor 0) se a hipótese alternativa (valor 1). Em relação à Tabela 8.2 verifica-se que as séries residuais que indiciavam não normalidade, depois de filtrados pelo modelo obtido, passaram a não indiciar não normalidade, que é o caso das de entrada LBC4, LBC8 TemC4 e TemC5.

Comparativamente, o desvio padrão da série de saída passou de 48.3 para 35.3 que poderá corresponder a uma redução da variabilidade do processo no máximo de 27%.

Para se ter uma melhor percepção da melhoria em termos da normalidade dos dados construíram-se os gráficos da distribuição empírica cumulativa (Figura 8-7) das séries de saída do modelo zero e do modelo obtido, em que é claramente visível uma maior concordância com a distribuição cumulativa normal para o segundo caso.

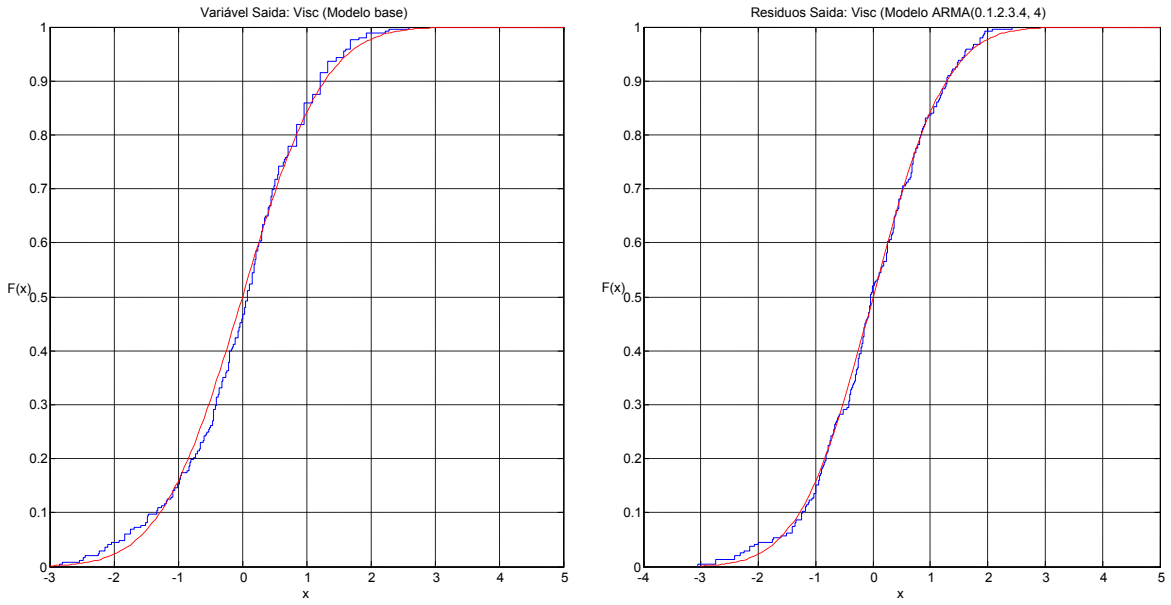


Figura 8-7 Gráfico da distribuição empírica cumulativa das variáveis de saída do modelo zero e do modelo ARMA(0.1.2.3.4, 4)

Ainda no sentido de se avaliar qualitativamente o modelo obtido, simulou-se a resposta do modelo às séries de entrada dos dados originais (Figura 8-8).

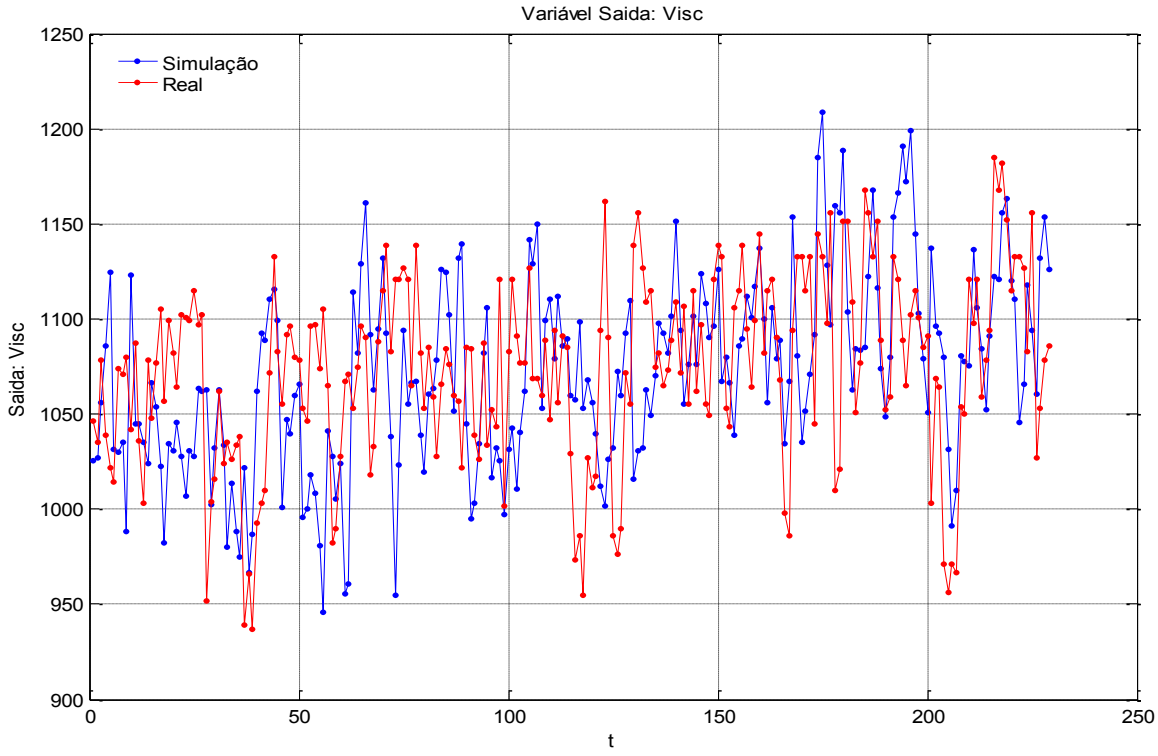


Figura 8-8 Resultados simulados versus resultados reais

8.4 Integração do controlo estatístico com engenharia de controlo

Antes de se integrar o controlo estatístico com a engenharia de controlo é conveniente que o sistema de controlo seja eficaz e robusto. Deste modo, numa primeira fase implementou-se e testou-se o algoritmo de controlo ao nível de simulação, e só após obter resultados satisfatórios a este é que se passou para a fase de integração do controlo estatístico. Cabe aqui dizer que a forma de sintonização dos parâmetros inerentes ao algoritmo de controlo foi feita de forma empírica por forma a queimar etapas e a obter mais rapidamente uma solução aceitável que demonstrasse a viabilidade da metodologia.

8.4.1 Estratégia de Controlo

A estratégia básica de controlo passa por um sistema de controlo por realimentação, ou malha fechada ou ainda anel fechado (*feedback/closed loop*). A estrutura base do modelo de controlo utilizado é a que é desenvolvida a partir do diagrama de blocos da Figura 6-3 com alguma alterações inerentes ao ajustamento ao modelo (Figura 8-9).

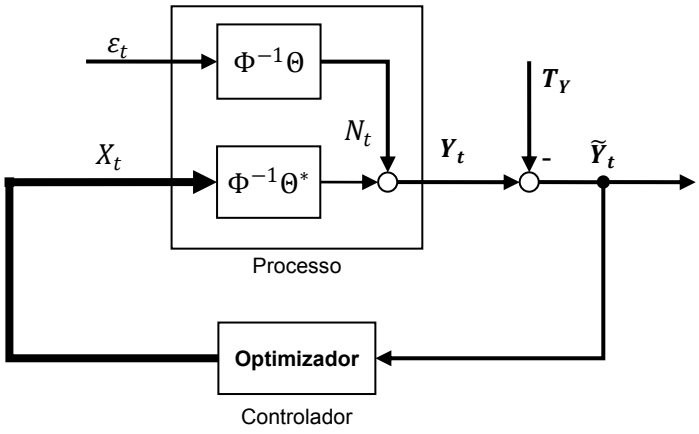


Figura 8-9 Diagrama de blocos base do sistema de controlo

O Processo é descrito pelo modelo ARMA(0.1.2.3.4, 4) desenvolvido no ponto 8.3.2 dado pela expressão **(8.4**) e cujo coeficientes se apresentam no ANEXO VI. O vector de entrada X_t é constituído pelas seis variáveis de entrada (AA, IS, DLBC4, DLBC8, TemC4 e TemC5) e a variável $\boldsymbol{Y_t}$ é composta pela saída do sistema (Viscosidade) que se pretende que se mantenha a um determinado nível $\boldsymbol{T_Y}$.

8.4.1.1 Controlador

De acordo com o ponto 6.2.3.1, o controlador utiliza uma derivação da versão multivariada do optimizador de Clarke e Gawthrop idêntica à utilizada por (Del Castillo & Yeh, An adaptive run-to-run optimizing controller for linear and nonlinear semiconductor processes, 1998) e já apresentado na expressão (7.10) dado por

$$\boldsymbol{J} = \left[\left(\hat{\boldsymbol{y}}_{\boldsymbol{t+b+1|t}} - \boldsymbol{T}\right)'\boldsymbol{W}\left(\hat{\boldsymbol{y}}_{\boldsymbol{t+b+1|t}} - \boldsymbol{T}\right) + (\boldsymbol{x_t} - \boldsymbol{x_{t-1}})\boldsymbol{\Gamma}(\boldsymbol{x_t} - \boldsymbol{x_{t-1}})\right] \quad (8.5)$$

Que é minimizado sujeito às restrições na entrada e na saída:

$$\boldsymbol{L_x \leq x_t \leq U_x} \quad (8.6)$$

$$\boldsymbol{L_y \leq y_t \leq U_y} \quad (8.7)$$

Aqui, em principio $\boldsymbol{\hat{y}_{t+b|t}}$ seria a previsão $(b+1)$ períodos à frente de $\boldsymbol{Y_{t+b+1}}$ condicional à informação existente no instante t. $\boldsymbol{T}$ é o vector das referências das variáveis de saída, $\boldsymbol{W}$ é uma matriz diagonal com as ponderações (prioridades) das saídas w_i e Γ é uma matriz diagonal com as ponderações (custos) das entradas. O vectores $\boldsymbol{L_x}$, $\boldsymbol{U_x}$, $\boldsymbol{L_y}$ e $\boldsymbol{U_y}$ estabelecem o domínio e contradomínio de actuação do processo. Isto é, os valores das entradas estão constrangidos à zona entre $\boldsymbol{L_x}$ e $\boldsymbol{U_x}$, em que $\boldsymbol{L_x}$ é o limite inferior e $\boldsymbol{U_x}$ é o limite superior; enquanto que as respostas do processo não devem ultrapassar os limites inferiores $\boldsymbol{L_y}$ e os limites inferiores $\boldsymbol{U_y}$ A optimização de J em ordem aos factores controláveis $\boldsymbol{x_t}$ é realizada com recurso ao cálculo numérico através ao método "*Penalty-Barrier*" (Jen, Jiang, & Fan, 2004), (Chen & Goldfarb, 2006). Neste caso recorreu-se à função do MATLAB *fmincon*.

8.4.1.2 Modelo Para o controlador

Esta estratégia de controlo insere-se dentro da categoria do controlo óptimo, pelo que, para implementar esta estratégia de controlo, o controlador necessita do modelo do processo, ou seja, o controlador necessita do modelo para poder prever a saída do processo $(b+1)$ períodos à frente ($\boldsymbol{Y_{t+b+1}}$). Uma vez que, neste caso, o objectivo é simular o processo e a respectiva resposta do controlador, não faz sentido o processo e o controlador terem o mesmo modelo, até porque normalmente o controlador é apenas uma linearização do comportamento do processo em redor de um ponto alvo de actuação, como tem sido referido ao longo deste texto. Por estas razões e baseado no estudo feito na secção 8.3, optou-se por um modelo autoregressivo puro de segunda ordem com a seguinte estrutura

$$\boldsymbol{(I_k - \Phi_0)Z_t = \Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + \varepsilon_t} \quad (8.8)$$

em que, como anteriormente, $\boldsymbol{Z_t} = [X_t^{'},\ Y_t^{'}]^{'}$, o que implica que, para este processo, os elementos da matriz Φ_0 serão nulos, à excepção dos seis primeiros elementos da sétima linha. Seguindo a notação introduzida, este modelo foi denominado por AR(0.1.2). Os resultados globais do ajustamento deste modelo podem ser consultados no ANEXO VII. Abaixo apresentam-se alguns indicadores.

Tabela 8.14 Indicadores de ajustamento do modelo AR(0.1.2)

$\lvert\tilde{\Sigma}\rvert$	R^2	Q_{10}	*P-value* (Q_{10})	*AIC*	*HQ*
6513	0,4280	537,36	1	9,977	10,82

$$\tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} 7.8729 & 0.9729 & -1.8418 & 0.2044 & -0.0250 & -0.1070 & -7.1655 \\ 0.9729 & 2.1463 & -0.3751 & 0.0296 & 0.0713 & 0.0248 & -1.7054 \\ -1.8418 & -0.3751 & 1.6555 & -0.1333 & -0.1061 & 0.0498 & -0.2663 \\ 0.2044 & 0.0296 & -0.1333 & 0.1053 & 0.0082 & 0.0394 & -0.2663 \\ -0.0250 & 0.0713 & -0.1061 & 0.0082 & 0.9926 & 0.6178 & -0.3330 \\ -0.1070 & 0.0248 & 0.0498 & 0.0394 & 0.6178 & 3.0954 & -3.0466 \\ -7.1655 & -1.7054 & -0.2663 & -0.8022 & -0.3330 & -3.0466 & 1379.80 \end{bmatrix}$$

$$\tilde{V} = \begin{bmatrix} 2.8059 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.3467 & 1.4234 & 0 & 0 & 0 & 0 & 0 \\ -0.6564 & -0.1036 & 1.1018 & 0 & 0 & 0 & 0 \\ 0.0728 & 0.0031 & -0.0773 & 0.3065 & 0 & 0 & 0 \\ -0.0089 & 0.0522 & -0.0967 & 0.0039 & 0.9902 & 0 & 0 \\ -0.0381 & 0.0267 & 0.0250 & 0.1435 & 0.6241 & 1.6378 & 0 \\ -2.5537 & -0.5761 & -1.8174 & -2.4629 & -0.4966 & -1.4775 & 36.893 \end{bmatrix}$$

Em relação ao modelo ARMA(0.1.2.3.4, 4), este modelo não consegue eliminar os indícios de correlação cruzada (marginal) existente na *lag* 4 (Figura 8-10) e apresenta um valor do determinante da matriz de covariância residual muito mais elevado, mas consegue valores aproximados do desvio padrão da variável de saída e de $\boldsymbol{R^2}$ muito semelhantes com uma redução significativa do numero de parâmetros, conseguindo-se assim baixar também os indicadores AIC e HQ.

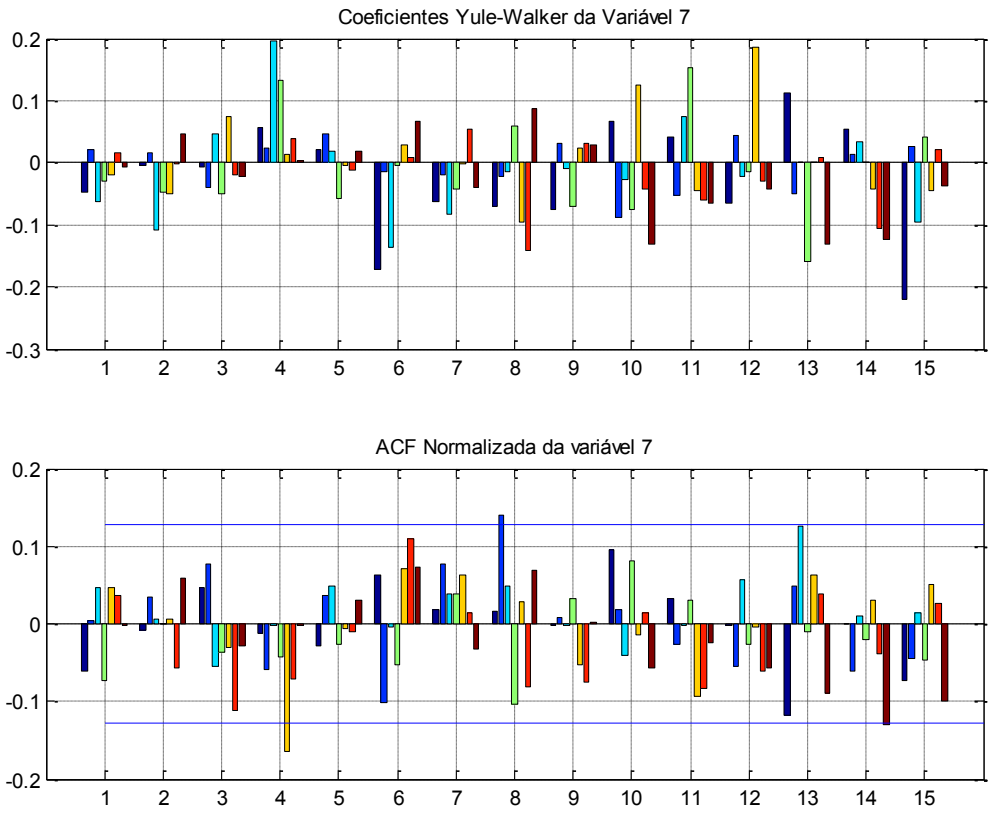


Figura 8-10 Coeficientes de Yule-Walker e ACF dos resíduos da variável de saída para o modelo AR(0.1.2)

8.4.1.3 Implementação do modelo

A partir do modelo determinado no ponto anterior pode-se então retirar a função de transferência a introduzir no controlador que será dada por

$$\tilde{y}_t = -2.364x_{3_t} + 4.368x_{2_{t-1}} - 2.841x_{5_{t-1}} - 2.886x_{6_{t-1}} + 1.443x_{1_{t-2}} + 3.589x_{6_{t-2}} + 0.465y_{t-1} + \varepsilon_{y_t} \quad (8.9)$$

ou

$$\tilde{y}_t = -2.364 \cdot DLBC4_t + 4.368 \cdot IS_{t-1} - 2.841 \cdot TemC4_{t-1} - 2.886 \cdot TemC5_{t-1} + 1.443 \cdot AA_{t-2} + 3.589 \cdot TemC5_{t-2} + 0.465 \cdot y_{t-1} + \varepsilon_{y_t} \quad (8.10)$$

em que $\tilde{y}_t$ é a diferença para a média (μ). Ao analisar esta expressão e a expressão do controlador (8.5) surge um problema que até agora ainda não foi mencionado, talvez porque só surge nas situações multivariadas, uma vez que temos variáveis de entrada desfasadas de uma *lag* e variáveis de entrada desfasadas de duas *lags*. A questão que se coloca é a seguinte: Qual o valor de *b* na expressão (8.5)?

A implementação em espaço de estados terá possivelmente resposta (?) para esta pergunta, mas essa abordagem não foi feita neste trabalho. A solução adoptada neste trabalho, que só foi possível devido ao facto de se desenvolver a metodologia para sistema MIMO, foi a de tentar minimizar a saída do processo simultaneamente um passo à frente e dois passos à frente, ou seja, a variável $\hat{y}_{t+b+1|t}$ que até aqui era uma escalar, passa a ser tratada como uma variável vectorial definida por $\left[\hat{y}_{t+1|t},\ \hat{y}_{t+2|t}\right]'$ em que cada uma das componentes são definidas por

$$\hat{y}_{t+1|t} = \mu - 2.364x_{3_{t+1}} + 4.368x_{2_t} - 2.841x_{5_t} - 2.886x_{6_t} + 1.443x_{1_{t-1}} + 3.589x_{6_{t-1}} + 0.465y_t$$

$$\hat{y}_{t+2|t} = \mu - 2.364\hat{x}_{3_{t+2}} + 4.368\hat{x}_{2_{t+1}} - 2.841x_{5_{t+1}} - 2.886x_{6_{t+1}} + 1.443x_{1_t} + 3.589x_{6_t} + 0.465\hat{y}_{t+1} \quad (8.11)$$

Ou seja, em cada iteração existem variáveis de entrada que são estimadas simultaneamente um passo à frente e dois passos à frente, mas obviamente, no sistema entram apenas varáveis referentes a esse instante.

Outro pormenor que se verifica na expressão (8.10) é a ausência da variável de entrada *DLBC8* (ou x_4). Como já foi referido, a ausência dessa variável não significa que ela não tenha influência na resposta do processo, significa, isso sim, que a variação dessa variável dentro dos valores admissíveis, ou seja, a gama de valores utilizados para identificação do processo, não altera significativamente a resposta do processo.

Embora na determinação do modelo do processo se tenha admitido a existência de alguma dinâmica nas variáveis de controlo, para efeitos de controlo, na ausência de mais informação do processo, considera-se as variáveis de controlo como determinísticas e instantâneas (o valor da constante de tempo é insignificante relativamente à constante de tempo do processo). Contudo, em relação às variáveis de controlo (entrada), o modelo intrínseco do processo separa a parte proveniente da dinâmica da parte proveniente do ajustamento devido à presença da componente média móvel presente no modelo.

Parâmetros
Para implementação do sistema de controlo tem que se determinar ainda os valores dos parâmetros W e Γ, os vectores referentes limites de operação $\boldsymbol{L_x}$, $\boldsymbol{U_x}$, $\boldsymbol{L_y}$ e $\boldsymbol{U_y}$ e ainda o valor de referência da saída do processo T. Estes parâmetros, normalmente, são definidos de acordo com os gestores dos processos tendo em conta condições económicas, operacionais e as especificações das características da saída do processo.

Um algoritmo especifico para sintonização dos parâmetros W e Γ poderia passar, por exemplo por simulação iterativa através de uma gama de valores possíveis dos parâmetros, e verificar quais os que produziam os resultados mais próximos dos pretendidos. Neste caso não se aprofundou o método, pelo que os valores utilizados foram determinados através de simulação, mas apenas até se obterem uma combinação de valores que satisfizessem as pretensões presentes. Um método mais preciso deverá ser implementado posteriormente.

Para efeitos de simulação, para os limites operacionais $\boldsymbol{L_x}$ e $\boldsymbol{U_x}$ das variáveis de entradas foram utilizados os mínimos e os máximos respectivamente, de cada série. Para o valor de referência da variável de saída T, numa primeira fase pensou-se em utilizar a média do processo, mas para verificar a eficiência do controlador optou-se por utilizar um valor diferente.

De notar que o controlador deverá estar projectado para admitir um valor de referência variável ($T(t)$, mas essa situação não é muito vulgar neste enquadramento.

Simulação do sistema de controlo
Para efeitos de validação da estratégia de controlo definida simulou-se o processo durante 600 observações com os seguintes parâmetros:

$$\boldsymbol{U_y = 1300}$$

$$\boldsymbol{L_y = 950}$$

$$\boldsymbol{W = I_2 16000}$$

$$\boldsymbol{\Gamma = I_9 0.0005}$$

Das observações simuladas obteve-se um desvio médio de **-3.9592** em relação ao valor de referência e um desvio padrão de **38.253.** A Figura 8-11 mostra o gráfico da variável de saída das 600 observações resultante da simulação do processo com controlo integrado. Os gráficos relativos às variáveis de entrada podem ser consultados no ANEXO VIII.

Comparativamente ao sistema sem controlo, para além de um maior controlo sobre o valor médio da resposta (parâmetro central), verificou-se um decréscimo moderado da variabilidade do processo, ou seja passou-se de 48.3 (Tabela 8.1) para aproximadamente 38.3. O desvio médio verificado poderá dever-se à diferença existente entre o modelo do processo e o modelo usado no controlador. Esse facto, de certo modo, reflecte o que se passa na realidade, uma vez que os modelos obtidos nunca conseguem captar a totalidade do comportamento dos processos que normalmente são de ordem bastante superior à admitida e frequentemente tem comportamentos não lineares que são captados apenas parcialmente pelos modelos lineares. Ou seja, os modelos nunca são perfeitos. O desvio

médio verificado, também apelidado de erro de offset, pode no entanto ser compensado com a introdução de um compensador PI ou apenas I.

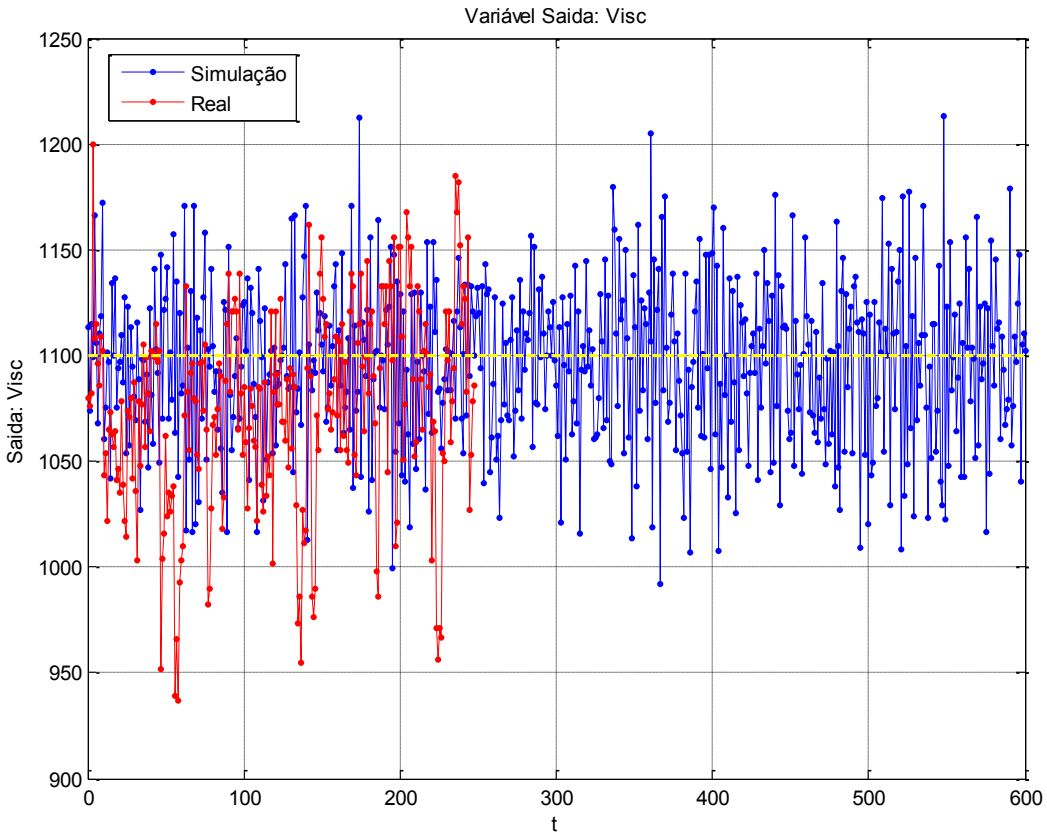


Figura 8-11 Gráfico da variável de saída do sistema com controlo

8.4.2 Estratégia de integração

A integração do controlo estatístico com engenharia de controlo pode ter vários objectivos, mas nem sempre aplicáveis.

Um dos objectivos poderá ser a redução da variabilidade dos parâmetros de entrada com a introdução de cartas, normalmente EWMA ou CUSUM de controlo à saída do processo com a função de delimitar (*treshold*) a zona morta, sendo que o controlador só actua quando a carta identifica uma situação de "fora de controlo" (ver Figura 7-2).

Outro dos objectivos poderá ser a implementação do chamado SPC algorítmico ou ASPC. Neste caso o controlo estatístico tem a sua função normal de detectar causas especiais que conduzam à melhoria do processo, que não seriam possíveis de detectar devido ao facto do controlador compensar possíveis alterações no processo mantendo a resposta dentro dos limites de controlo. Esta estratégia passa por implementar cartas de controlo à entrada e saída do processo (Figura 8-12). Se à saída for detectada uma situação fora de controlo será porque o controlador não conseguiu compensar a alteração verificada no processo. Se o controlador conseguir compensar a alteração, então esta deverá ser detectada nas cartas

de controlo implementadas na entrada do processo devido à compensação "fora dos limites" exigida devido às alterações do processo.

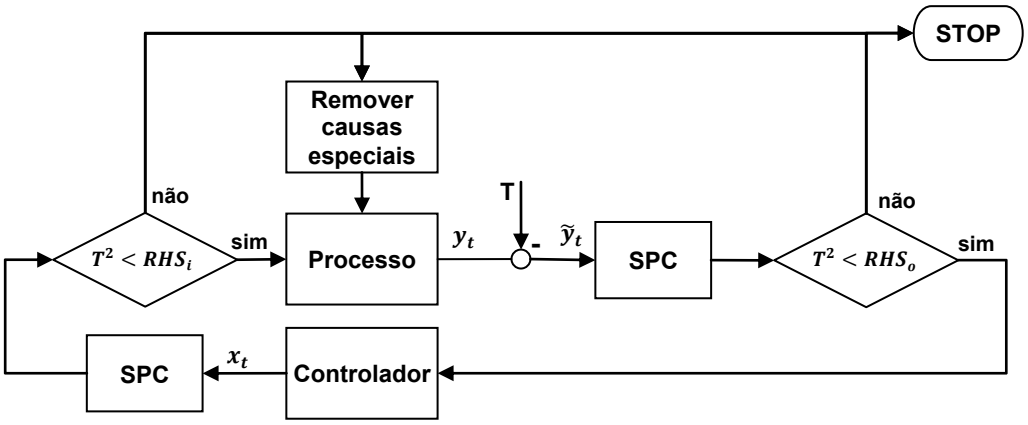


Figura 8-12 Diagrama de blocos de possível implementação do SPC algorítmico (ASPC)

Quanto à primeira abordagem (zona morta), verificou-se que, no processo em causa, a viabilidade era reduzida devido ao reduzido ratio entre o desvio padrão mínimo (desvio padrão com o processo ajustado em todos os instantes – sem zona morta) e o desvio padrão sem qualquer controlo. No que diz respeito à segunda abordagem (ASPC), conclui-se que, devido ao sistema de controlo implementado, não era fácil detectar qualquer situação "fora de controlo" nas variáveis de controlo (entrada) devido ao facto de estas estarem constrangidas (limitadas entre um máximo e um mínimo) pelo próprio controlador, pelo que a verificar-se qualquer situação "fora de controlo", esta teria que ser nas variáveis de saída.

8.4.2.1 Modelo

Uma vez postas de parte estas duas abordagens, optou-se pelo esquema mais generalista apresentado na Figura 7-1 em que o controlo estatístico tem a dupla função de alterar os parâmetros do controlador, caso seja viável, e simultaneamente detectar causas especiais.

No modelo implementado (

Figura 8-13), utiliza-se uma carta EWMA para monitorar a média do processo. Sempre que é detectada uma alteração da média do processo $(|\boldsymbol{z_t}| > L\sigma)$, o sistema verifica se o controlador consegue compensar essa alteração, ou seja, se $\mu_{t+1} = \mu_t + \delta$ estiver compreendido entre o valor mínimo e o valor máximo admissível pelo controlador, respectivamente μ_L e μ_U. Se o controlador consegue compensar a alteração então é lançado um alerta, o controlador actualiza o parâmetro que foi alterado (neste caso a média), o processo continua, mas espera-se que a causa seja identificada se for o caso (poderá tratar-se de um processo com média oscilante) e removida. Caso contrário, o processo pára para que seja identificada a causa e removida. Ou seja, perante uma alteração no processo, é sempre emitido um alerta, podendo no entanto, o processo parar ou não.

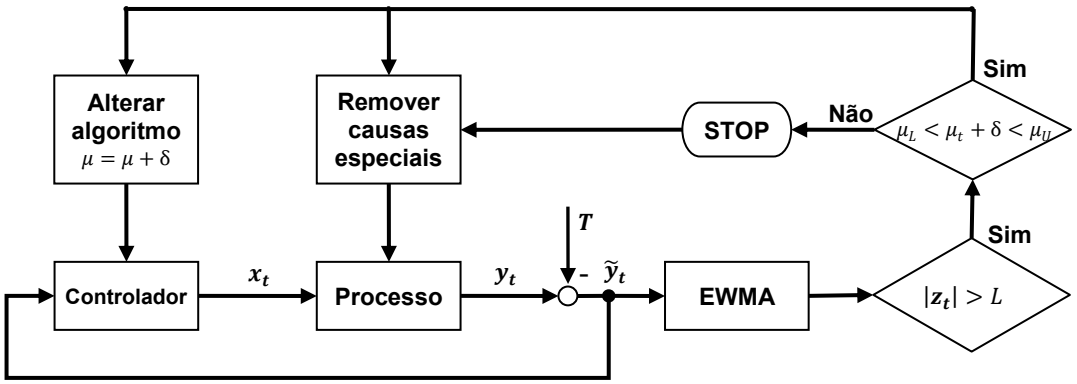


Figura 8-13 Integração EPC/SPC (Monitorização da média do processo)

8.4.2.2 Carta EWMA

Como o processo analisado apenas tem uma saída (sistema MISO) usou-se uma carta EWMA univariada (Montgomery D. C., 2001) para monitorizar a variável de saída. O valor de z_t (
Figura 8-13) é determinado pela expressão

$$\boldsymbol{z_t = \lambda y_t + (1-\lambda) z_{t-1}}$$

E os limites de controlo dados pelas expressões assimptóticas

$$\boldsymbol{UCL = \mu_0 + L\sigma \sqrt{\frac{\lambda}{(2-\lambda)}}}$$

$$\boldsymbol{LCL = \mu_0 - L\sigma \sqrt{\frac{\lambda}{(2-\lambda)}}}$$

De notar que, se a resposta do processo fosse multivariável, provavelmente ter-se-ia que utilizar uma carta EWMA para cada variável, uma vez que a carta EWMA multivariada (Lowry, Woodall, Champ, & Rigdon, 1992) não informa qual a variável que se alterou, pelo que não seria possível detectar qual o parâmetro do controlador a actualizar. Nesse caso, contudo, seria conveniente retirar a componente de correlação das variáveis antes do processo de monitorização (ver ponto 6.1.1.2). Comparativamente ao modelo da Figura 7-4 (Del Castillo & Yeh, 1998), que utiliza uma carta EWMA multivariável a diferença é que nesse caso, sempre que se verifica um situação de "fora de controlo", os parâmetros do modelo são todos actualizados, pelo que não existe a necessidade de verificar qual o parâmetro que se alterou.

8.4.2.3 Simulação

No campo da simulação o trabalho efectuado foi relativamente reduzido, e feito apenas a título exemplificativo do que poderá ser o ponto de partida para trabalhos futuros na área da integração de EPC/SPC.

Com base no modelo da Figura 8-13 foram simulados três cenários diferentes. Em todos os casos foram usados os seguintes parâmetros para o controlador:

$$\boldsymbol{U_y = 1300}$$

$$\boldsymbol{L_y = 950}$$

$$\boldsymbol{W = I_2 10000}$$

$$\boldsymbol{\Gamma = I_9 0.0005}$$

No primeiro cenário tentou-se simular alterações de um (1) desvio padrão (σ) na média do processo e verificar o comportamento da resposta da carta EWMA, do controlador e da resposta do processo. Como o objectivo não passava pela paragem do processo, os limites μ_L e μ_U foram estabelecidos de forma a não serem atingidos. Neste cenário foram utilizados os seguintes valores para a carta EWMA (Montgomery D. C., 2001)

$$\boldsymbol{\lambda = 0.2}$$

$$\boldsymbol{L = 2.962}$$

Que para uma alteração de um (1) desvio padrão (σ) na média do processo corresponde a um ARL (Average Run Lenght) de 10,5.

Num segundo cenário, simularam-se alterações de meio (0.5) desvio padrão (σ) na média do processo, sendo admissíveis variações entre $\mu_0 - 1.5\sigma$ e $\mu_0 + 1.5\sigma$ sem que o processo fosse parado devido a alarme de alteração no processo não compensada pelo controlador. Também aqui o objectivo não passa por parar o processo, mas sim verificar o comportamento do sistema integrado com detecção de amplitudes mais baixas na média. Neste cenário foram utilizados os valores de $\boldsymbol{\lambda = 0.4}$ e $\boldsymbol{L = 3.054}$ que corresponde a um ARL de 71.2 para 0.5σ e de 14.3 para 1σ.

Num terceiro cenário, simularam-se alterações de meio (0.5) desvio padrão (σ) na média do processo, sendo admissíveis variações entre $\mu_0 - 1.5\sigma$ e $\mu_0 + 1.5\sigma$ sem que o processo fosse parado devido a alarme de alteração no processo não compensada pelo controlador. Também aqui o objectivo não passa por parar o processo, mas sim verificar o comportamento do sistema integrado com detecção de amplitudes mais baixas na média. Também neste cenário foram utilizados os valores de $\boldsymbol{\lambda = 0.2}$ e $\boldsymbol{L = 2.962}$ que correspondem a um ARL de 41.8 para 0.5σ e de 10.5 para 1σ.

Em todos os cenários a simulação foi composta por 500 observações e foi introduzida uma alteração de um (1) desvio padrão (σ) na média do processo entre as observações 100 e 300.

8.4.2.4 Resultados

Cenário 1

Das observações simuladas sob as condições descritas para o cenário 1, obteve-se um desvio médio de -**1.0729** em relação ao valor de referência e um desvio padrão de **37.242**. A Figura 8-14 mostra o gráfico da variável de saída das 500 observações resultante da

simulação do processo com integração SPC/EPC. Os gráficos relativos às variáveis de entrada podem ser consultados no ANEXO IX. Como se pode verificar, a rápida detecção da carta EWMA e a consequente alteração do modelo interno do controlador impediu que a perturbação introduzida no processo influenciasse significativamente a saída do processo.

A Figura 8-15 mostra a carta EWMA, com os respectivos limites, resultante da simulação do processo nas condições estabelecidas para o cenário 1. A Tabela 8.16 mostra alguns dos pontos mais significativos dessa carta. No processo verificou-se uma alteração na média de um (1) desvio padrão (σ) no instante 100. Como se pode verificar, a carta detectou uma alteração na média logo na observação 108 (o ARL nestas condições é de 10.5). Após a detecção, o sistema alterou o parâmetro da média adicionando-lhe 1σ como está estipulado, e o modelo interno do controlador ficou de acordo com o processo alterado no instante 100.

O aumento de 1σ na média do processo é retirado na observação 300. Esta alteração é detectada pela carta logo na observação 306 com a consequente actualização do respectivo parâmetro no modelo interno do controlador.

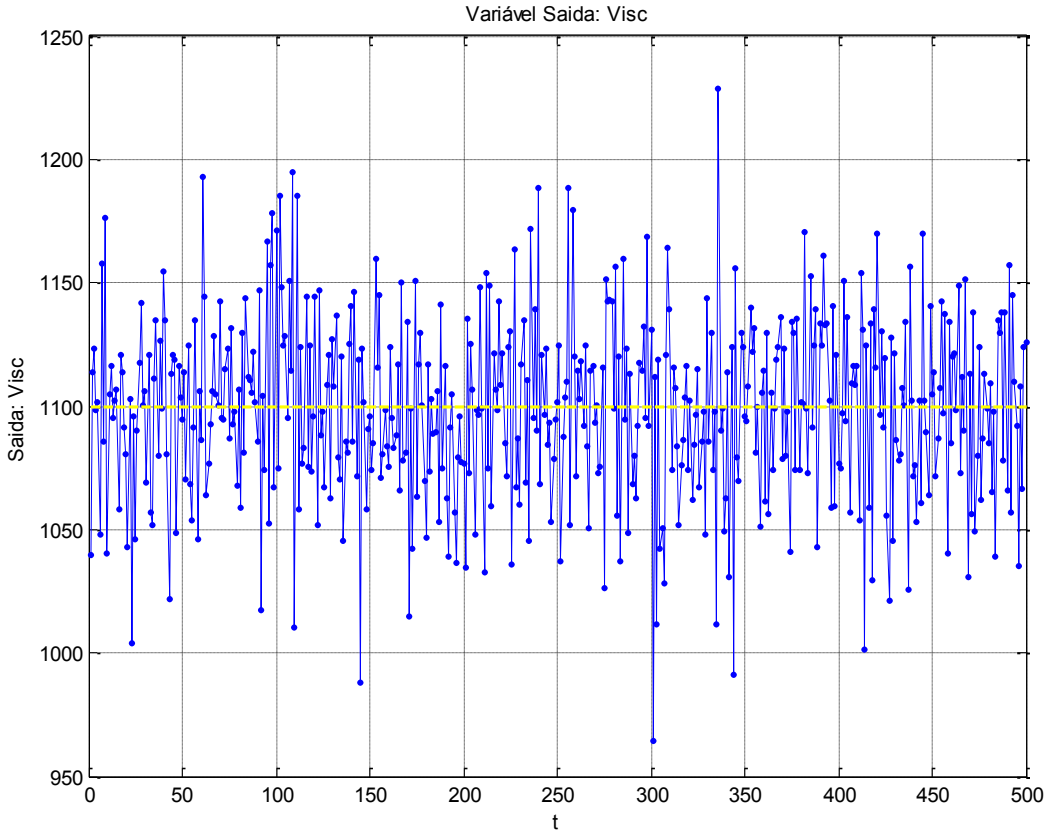


Figura 8-14 Resposta do processo (cenário 1)

Embora a alteração introduzida no processo não seja perceptível na saída do processo, as suas consequências são perfeitamente visíveis no comportamento das variáveis de entrada. Esta constatação vai ao encontro da teoria de suporte da metodologia ASPC (*algotithmic statistical process contro*l) que argumenta que se a perturbação não é detectável a partir da

resposta do processo quando este está sob controlo por realimentação, então ela será detectável através das alterações verificadas nas variáveis de entrada.

Tabela 8.15 Alguns pontos da carta EWMA

Observação	LIC	LSC	z_t	δ
1	60,440	-9,307	25,567	0,000
2	60,440	-9,307	30,234	0,000
107	60,440	-9,307	51,222	0,000
108	**60,440**	**-9,307**	**65,048**	**35,321**
305	60,440	-9,307	-4,231	35,321
306	**60,440**	**-9,307**	**-12,603**	**0,000**
500	60,440	-9,307	23,721	0,000

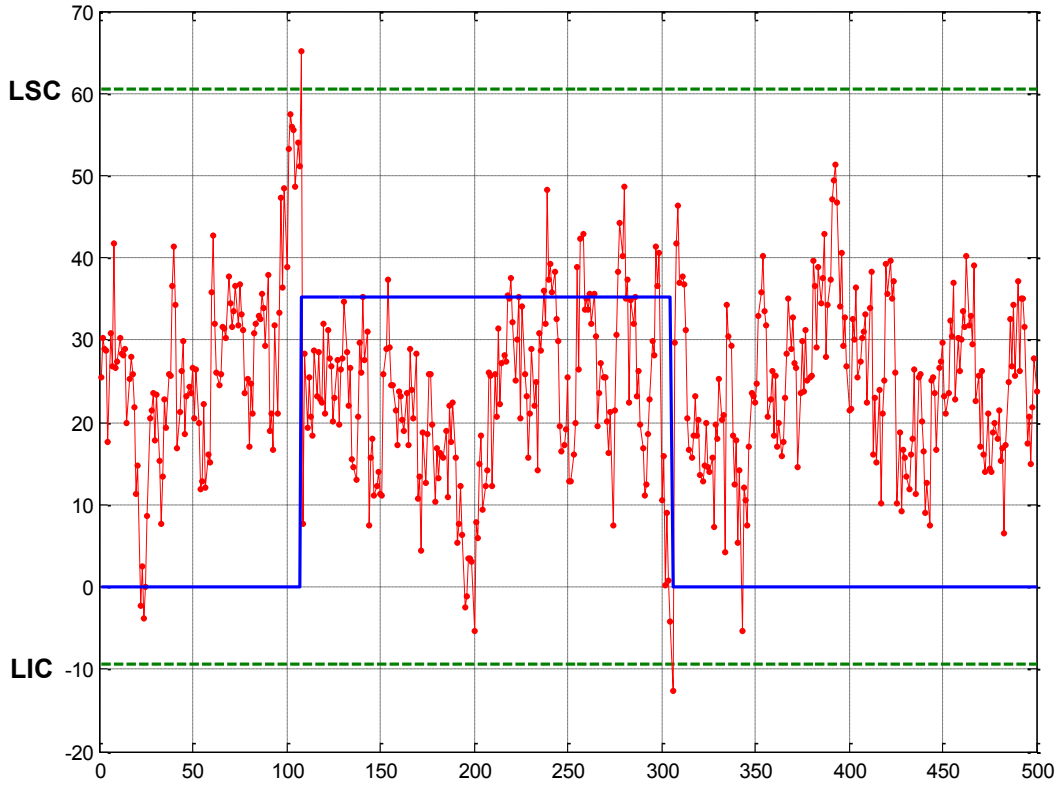


Figura 8-15 Carta EWMA (cenário 1)

De facto, na Figura 8-16 são perfeitamente visíveis as consequências das alterações provocadas no processo, com as variáveis de controlo a corresponder às solicitações do controlador no sentido de manter a resposta próxima dos valores de referência.

Há a salientar contudo que, neste caso, a detecção foi efectuada na variável de saída, e que em princípio, este facto deve despoletar uma análise ao processo no sentido de identificar a causa da alteração e proceder á sua remoção no sentido da melhoria contínua do processo. Existem no entanto processos que não têm média constante, ou seja a média oscila em torno de um "valor médio". Nesses casos, a implementação de sistema de controlo semelhante a este poderá diminuir significativamente a variabilidade do processo.

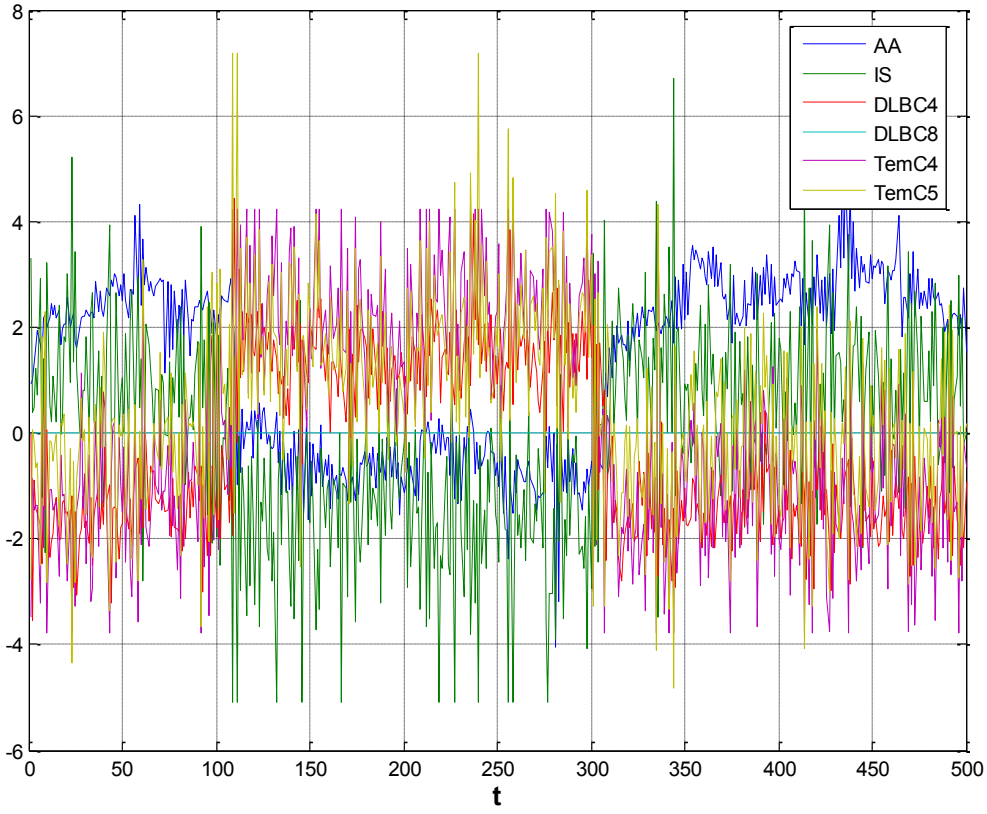


Figura 8-16 Variáveis de entrada do processo (cenário 1)

Cenário 2

Das observações simuladas sob as condições descritas para o cenário 2, obteve-se um desvio médio de **-6.0364** em relação ao valor de referência e um desvio padrão de **43.974.** A Figura 8-17 mostra o gráfico da variável de saída das 500 observações resultante da simulação do processo com integração SPC/EPC. Os gráficos relativos às variáveis de entrada podem ser consultados no ANEXO X.

A Figura 8-18 mostra a carta EWMA, com os respectivos limites, resultante da simulação do processo e a Tabela 8.16 mostra alguns dos pontos mais significativos dessa carta. No processo verificou-se uma alteração na média de um (1) desvio padrão (σ) no instante 100. Como se pode verificar, a carta detectou uma alteração na média na observação 110. Após a detecção, o sistema alterou o parâmetro da média adicionando-lhe 0.5σ como está estipulado. No instante 143 a carta detecta outra alteração na média (o ARL para 0.5σ é de 41.8) e o sistema actualiza novamente o respectivo parâmetro, ficando o parâmetro da média do modelo interno do controlador de acordo com a perturbação sofrida.

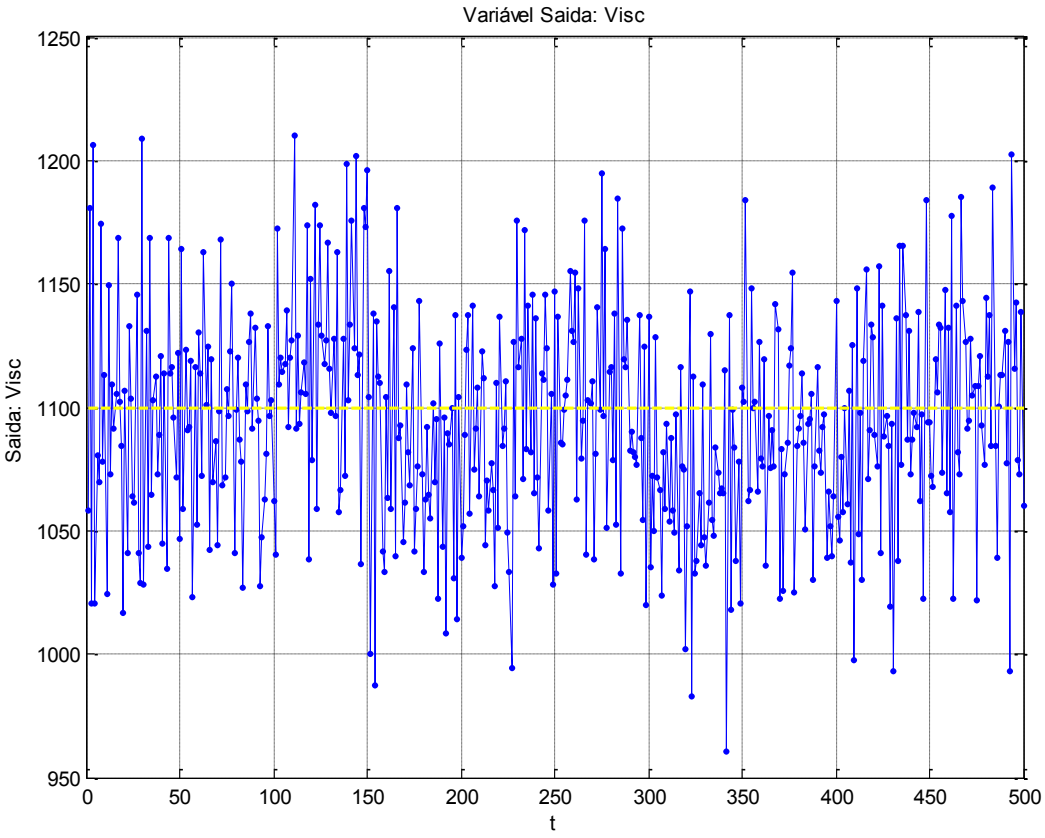


Figura 8-17 Saída do processo (cenário 2)

Tabela 8.16 Alguns pontos da carta EWMA

Observação	LIC	LSC	z_t	δ
1	79,501	-28,368	25,567	0,000
110	**79,501**	**-28,368**	**81,503**	**17,660**
142	**79,501**	**-28,368**	**62,825**	**17,660**
143	**79,501**	**-28,368**	**88,626**	**35,321**
148	79,501	-28,368	67,419	35,321
149	**79,501**	**-28,368**	**89,034**	**52,981**
226	**79,501**	**-28,368**	**-40,386**	**35,321**
340	79,501	-28,368	11,801	35,321
341	**79,501**	**-28,368**	**-38,506**	**17,660**
429	79,501	-28,368	1,047	17,660
430	**79,501**	**-28,368**	**-31,908**	**0,000**
500	79,501	-28,368	47,450	0,000

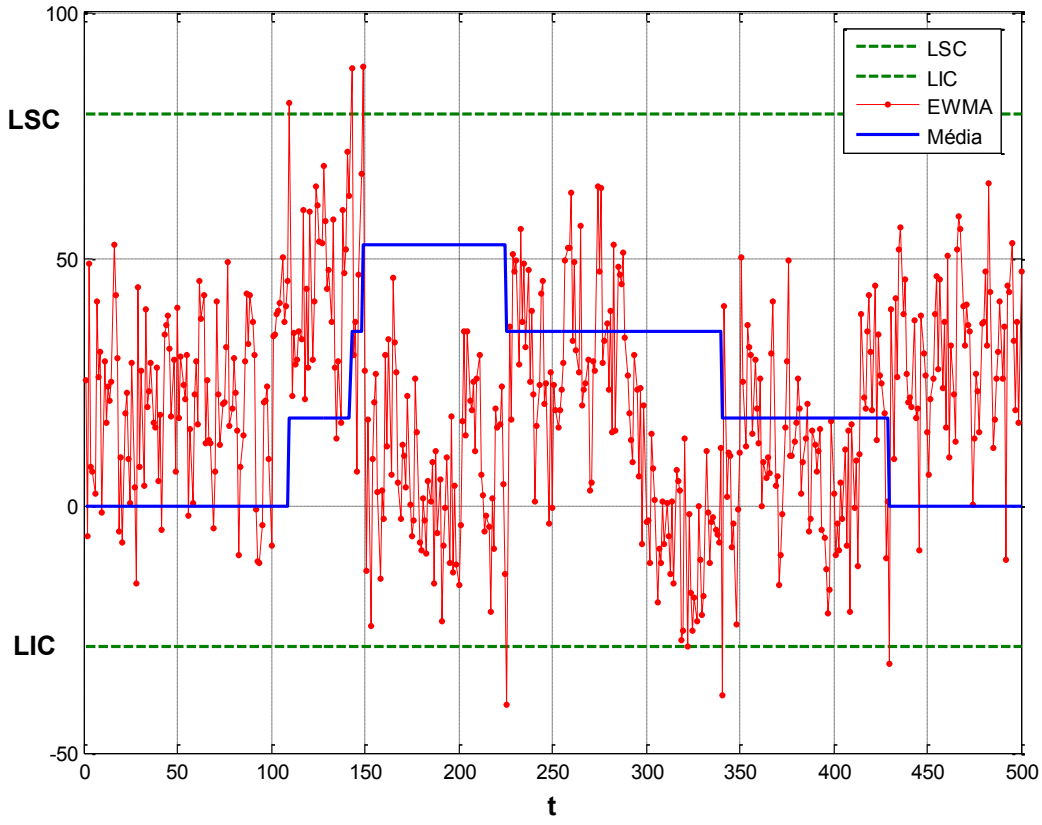


Figura 8-18 Carta EWMA (cenário 2)

Na observação 149, a carta detecta novamente um aumento da média do processo, mas neste caso trata-se de um falso alarme. Se fossem admissíveis apenas variações entre $\mu_0 - 1\sigma$ e $\mu_0 + 1\sigma$, o processo pararia para que fosse identificada a causa da variação. O falso alarme foi corrigido pelo próprio sistema no ponto 226, ficando o modelo interno do controlador novamente de acordo com a perturbação sofrida pelo processo.

O aumento de 1σ na média do processo é retirado na observação 300. Esta alteração só é detectada pela carta na observação 341 com a consequente actualização do respectivo parâmetro, mas olhando para a carta (Figura 8-18) verifica-se que esteve muito perto de ser detectada entre as observações 319 e 322. A alteração só é detectada totalmente (1σ) no ponto 430 passando então o modelo interno do controlador a estar novamente de acordo com o processo.

No gráfico referente à saída do processo (Figura 8-17) são perfeitamente identificáveis as zonas em que o modelo do controlador não está de acordo com o modelo do processo.

Na Figura 8-19 pode-se verificar as alterações efectuadas nas variáveis de entrada pelo controlador para compensar as perturbações detectadas. Nessa figura são perfeitamente detectáveis as restrições impostas pelo algoritmo de controlo, no que diz respeito aos limites inferiores e superiores impostos nas variáveis de entrada, nomeadamente nas variáveis TemC4 e IS. Esta situação é apelidada na literatura de controlo como “saturação dos actuadores”.

No global, salienta-se que apesar das perturbações impostas, o desvio médio em relação ao valor de referência e a própria variabilidade do processo situam-se em níveis aceitáveis, comparávelmente aos resultados reais (sem controlo nem perturbações).

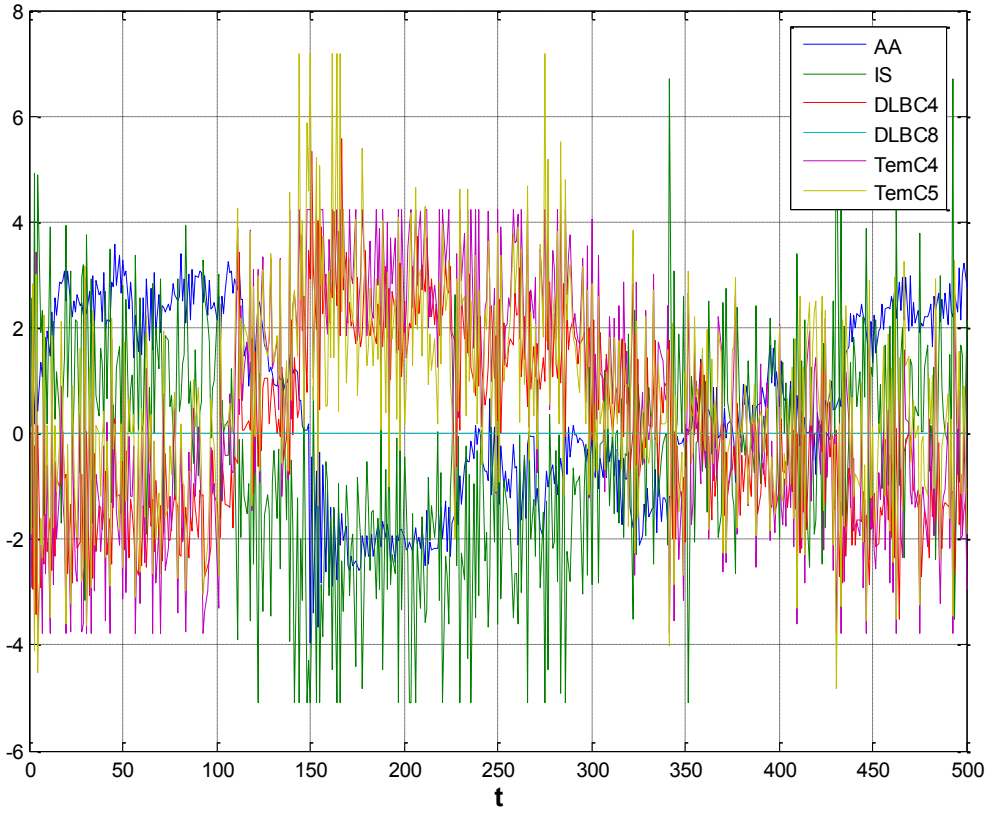


Figura 8-19 Variáveis de entrada do processo (cenário 2)

Cenário 3

Das observações simuladas sob as condições descritas para o cenário 3, obteve-se um desvio médio de **-2.4004** em relação ao valor de referência e um desvio padrão de **41.113.** A Figura 8-20 mostra o gráfico da variável de saída das 500 observações resultante da simulação do processo com integração SPC/EPC. Os gráficos relativos às variáveis de entrada podem ser consultados no ANEXO XI.

A Figura 8-21 mostra a carta EWMA, com os respectivos limites, resultante da simulação do processo e a Tabela 8.17 mostra alguns dos pontos mais significativos dessa carta. Tal como nos casos anteriores, no processo verificou-se uma alteração na média de um (1) desvio padrão (σ) no instante 100. Como se pode verificar, a carta detectou uma alteração na média na observação 105 (o ARL para 1σ é de 10.5). Após a detecção, o sistema alterou o parâmetro da média adicionando-lhe 0.5σ como está estipulado. No instante 115 a carta detecta outra alteração na média e o sistema actualiza novamente o respectivo parâmetro, ficando novamente de acordo com a perturbação sofrida.

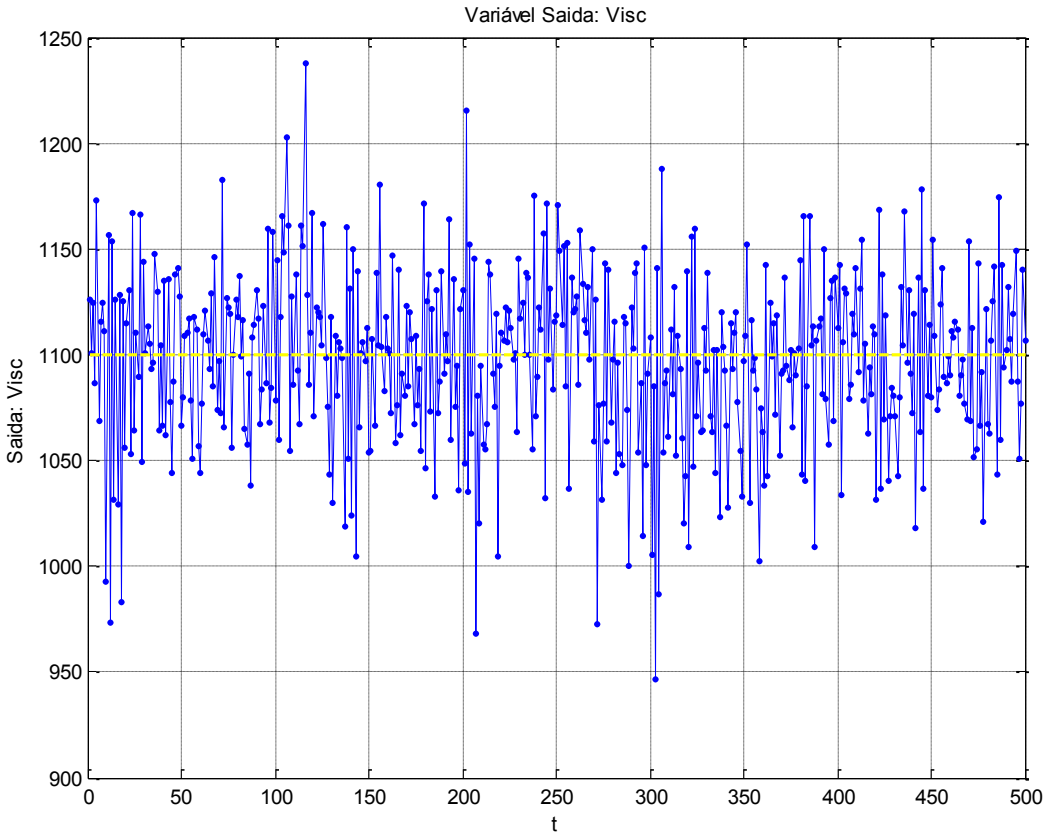


Figura 8-20 Saída do processo (cenário 3)

Tabela 8.17 Alguns pontos da carta EWMA

Observação	LIC	LSC	z_t	δ
1	60,440	-9,307	25,567	0,000
104	60,440	-9,307	48,634	0,000
105	**60,440**	**-9,307**	**64,565**	**17,660**
115	**60,440**	**-9,307**	**68,069**	**35,321**
301	60,440	-9,307	0,776	35,321
302	**60,440**	**-9,307**	**-24,990**	**17,660**
359	60,440	-9,307	-3,025	17,660
360	**60,440**	**-9,307**	**-9,678**	**0,000**
500	60,440	-9,307	18,671	0,000

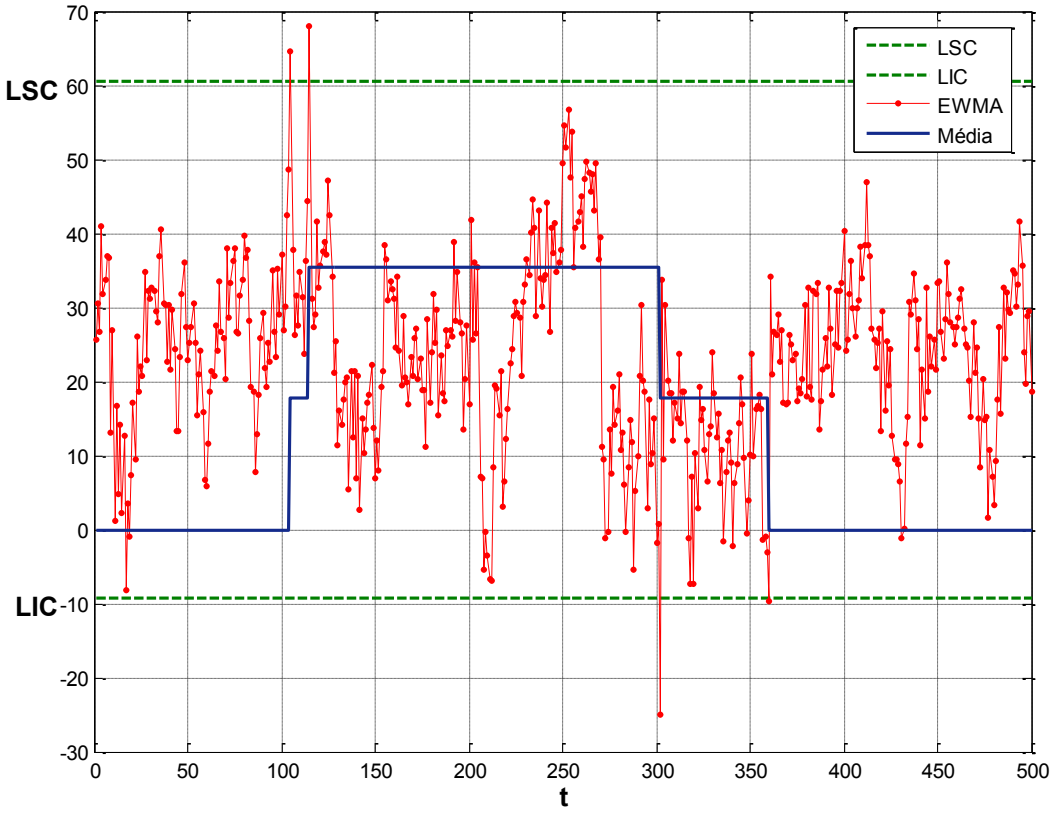


Figura 8-21 Carta EWMA (cenário 3)

O aumento de 1σ na média do processo é retirado na observação 300. Esta alteração é detectada pela carta logo na observação 302 com a consequente actualização do respectivo parâmetro. A alteração é detectada totalmente (1σ) no ponto 360 passando então o modelo interno do controlador a estar novamente de acordo com o processo.

No gráfico referente à saída do processo (Figura 8-20Figura 8-17) nota-se apenas uma ligeira subida na zona compreendida entre a observação 100 e 110, pouco perceptível, e uma ligeiro abaixamento na entre as observações 300 e 350. No global, são pouco perceptíveis as consequências das alterações que ocorreram.

No que diz respeito às variáveis de entrada, como se pode verificar na Figura 8-22, os efeitos provocados pelas acções de controlo nas variáveis de entrada são bem perceptíveis.

No global, verifica-se que, para além de um muito ligeiro aumento no parâmetro de dispersão, a resposta do processo praticamente não sentiu a perturbação imposta.

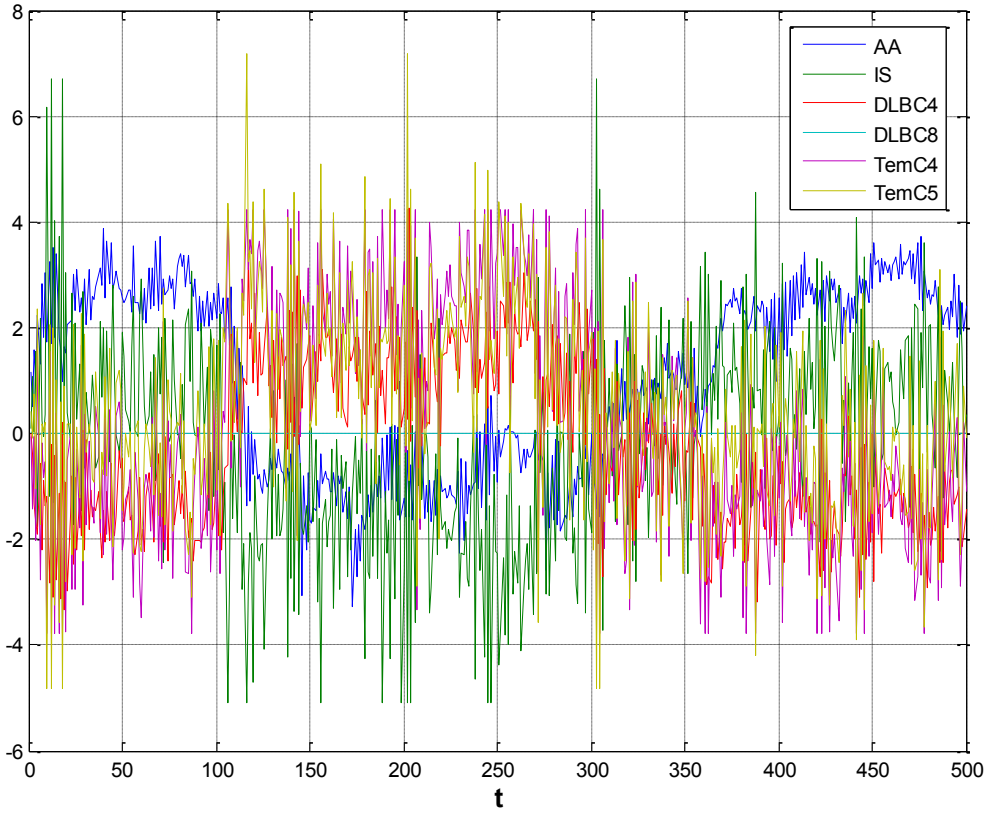


Figura 8-22 Variáveis de entrada do processo (cenário 3)

8.4.3 Conclusões

A primeira conclusão que se pode retirar é que as simulações efectuadas apontam no sentido de esta metodologia permitir controlar o parâmetro central da saída do processo, mantendo-o bastante próximo dos valores de referência (*target*), e simultaneamente reduzir a sua variabilidade mesmo perante a presença de alterações no processo.

Outra das conclusões que surge dos cenários simulados, é que se obtém melhores resultados com valores de lambda mais baixos. Em geral, verifica-se que valores no intervalo $0.05 \leq \lambda \leq 0.25$ conduzem a bons resultados na detecção de pequenas alterações (Montgomery D. C., 2001) e que valores de $L{\sim}3$ (limites de três sigma) funcionam razoavelmente bem para valores de $\lambda > 0.1$.

No seguimento do que se vem defendendo ao longo deste trabalho, a obtenção de um bom modelo, no sentido de representar correctamente o comportamento do processo, consiste numa ferramenta poderosíssima no sentido melhorar significativamente, ou mesmo optimizar o processo. Os métodos de simulação aqui apresentados são bastante básicos e com reduzida inferência estatística. O objectivo principal dos cenários simulados, basicamente foi demonstrar algumas das potencialidades da metodologia e exemplificar o trabalho que poderá ser desenvolvido a partir daqui.

Aqui simulou-se um caso prático duma estratégia de integração do controlo estatístico com engenharia de controlo no intuito de demonstrar as potencialidades da sinergia destas duas metodologias. Outras estratégias existem e foram aqui discutidas e outras existem que não foram abordadas neste trabalho. A estratégia a adoptar na implementação desta metodologia pode variar significativamente de implementação para implementação, de empresa para empresa, de processo para processo, daí que muitos autores nesta área, (Del Castillo, 2002) entre outros, argumentem que o reduzido sucesso verificado até agora na implementação desta metodologia está directamente ligado à dificuldade de a transformar num produto comercial genérico.

9 Conclusões, recomendações e trabalho futuro

O objectivo original deste trabalho apontava para o estudo e desenvolvimento de uma metodologia que permitisse integrar o conceito de controlo estatístico com o conceito de controlo do processo com aplicação prática em ambiente industrial. Esse objectivo foi sendo redefinido ao longo do próprio trabalho. Desde cedo se identificou a dificuldade de atingir esse objectivo, devido às condicionantes existentes quer em termos operacionais, existência de um caso de estudo, quer em termos de prazo de execução do trabalho. De salientar que, apesar de ser um tema bastante abordado recentemente nos meios académicos ligados à engenharia industrial, não se encontram ainda grandes exemplos de aplicação e implementação desta metodologia. Os exemplos mais estudados e mais mediáticos estão orientados para indústria de semicondutores e invariavelmente focados no mesmo processo.

Deste modo optou-se por redefinir os objectivos no sentido de se chegar tão longe quanto possível no desenvolvimento e implementação das ferramentas base que permitissem integrar e implementar os dois conceitos.

Também desde cedo se reconheceu que o principal pilar para a aplicação da metodologia passava pela identificação e modelação dos processos, no sentido de que quanto mais perto o modelo estiver da representação real do processo, ou seja, quanto mais fiável for o modelo, mais fácil se torna encontrar uma solução eficaz para o processo. Deste modo, grande parte do trabalho foi dedicado à identificação e modelação de processos dinâmicos através do estudo, desenvolvimento, implementação e validação de algoritmos que actualmente podem ser utilizados a partir de dados históricos de processos, sejam eles industriais ou não. Após atingido o objectivo principal, a disponibilidade restante foi canalizada para exemplificação do rumo a dar ao trabalho efectuado. Ou seja nos quatro grandes blocos que foram identificados como a composição das áreas de estudo inerentes a esta metodologia (identificação/modelação, controlo estatístico - SPC, engenharia de controlo - EPC e integração de EPC/SPC), optou-se por sedimentar razoavelmente a base onde assentam ou dois blocos a integrar.

Apesar de este trabalho estar orientado para processos gerais de entradas múltiplas e saídas múltiplas (MIMO), verifica-se que uma grande parte deste texto está dedicada a sistemas univariados (capítulos a 4 e 5). Esse facto está relacionado com a ideia que se interiorizou de que os conceitos básicos inerentes a esta matéria eram mais facilmente assimilados com modelos univariados, pelo que se optou por numa primeira fase fazer uma abordagem aos conceitos relacionados com sistemas univaridados e numa segunda fase generaliza-los para sistemas de entradas múltiplas e saídas múltipla.

Ainda no campo da identificação, apesar de grande parte deste trabalho se ter efectuado nesta área, ficou-se aquém do que se pretendia fazer, uma vez que metodologias de identificação como desenho de experiências, resposta em superfície, identificação em malha fechada ou identificação em espaço de estados, não foram abordadas ou foram abordadas de forma muito ligeira.

Apesar de a abordagem de representação em espaço de estados apenas se ter feito esporadicamente, cedo se concluiu que uma abordagem a sistemas multi-input multi-output (MIMO) deverá passar por esse tipo de representação uma vez que praticamente toda a filosofia por detrás da engenharia de controlo moderno, como seja o controlo óptimo, controlo preditivo, controlo adaptativo, etc. assenta nessa representação dos processos. No entanto, chegar a esse tipo de representação a partir do trabalho de identificação/modelação efectuado, é apenas mais um passo (Lütkepohl, 2007), ou seja, passar de uma representação com a estrutura (6.85) para uma representação do tipo (4.58)/(4.59) é óbvio. Nesta abordagem, contrariamente à abordagem de função de transferência têm-se em conta não só as variáveis de entrada mas também as chamadas variáveis de estado que representam o que se passa no interior do sistema. Olhando para o sistema em estudo pode-se facilmente concluir que, por exemplo, uma temperatura não deverá ser considerada uma variável de entrada, mas antes uma variável de estado que seria função de uma ou mais variáveis de entrada (um determinado fluxo de quantidade de calor e/ou energia, por exemplo). Nos sistemas de controlo baseados na representação de espaço de estados o controlo óptimo é feito através da realimentação das variáveis de estado, pelo que a optimização é sempre feita um passo à frente. Com isto pode-se concluir que, neste campo, a evolução natural deste trabalho deverá passar por uma abordagem mais profunda da representação deste processo em espaço de estados e a aplicação de metodologias de controlo preditivo mais avançadas que simultaneamente permitam assegurar um melhor controlo da resposta transitória.

Em relação ao segmento do controlo estatístico do processo, o desafio que se colocava era de que modo se conseguia implementar uma ou mais cartas de controlo de modo a que a informação retirada permitisse simultaneamente detectar alterações do processo e dar a informação ao controlador sobre os parâmetros a alterar. Entre as soluções avaliadas, CUSUM ou EWMA, optou-se por uma testar uma carta EWMA com resultados preliminares que se consideram satisfatórios. Como o caso concreto de estudo é um sistema MISO, e concluiu-se que não haveria necessidade de implementar cartas de controlo nas variáveis de entrada devido ao facto delas estarem constrangidas pelo controlador, não houve a necessidade de abordar sistemas de controlo estatístico multivariado. Como toda a abordagem do tema foi feita para sistemas de entradas múltiplas e saídas múltiplas, a questão que seguidamente se colocaria para um sistema MIMO é a seguinte: Para uma implementação multi-input / multi-output esquematizada pelo diagrama de blocos da Figura 8-13, qual o tipo de carta EWMA a utilizar?

Na resposta dada argumentou-se que teriam que ser utilizadas cartas univariadas, tantas quantas as variáveis de saída, mas as variáveis a monitorizar deveriam estar não correlacionadas (ver ponto 6.1.1.2) para evitar que uma alteração detectada numa das variáveis não induzisse um falso alarme noutra variável que com ela estivesse correlacionada. Esta posição continua a ser mantida, uma vez que se necessita de saber o mais rapidamente possível qual a variável que sofreu a alteração. Existe no entanto outra questão que se coloca em termos de ARL, ou seja, como é que detectada mais rapidamente uma situação de fora de controlo neste caso especifico, com cartas univariadas ou multivariadas?

A ferramenta identificada para esclarecer estas dúvidas passa pela simulação de uma série de modelos e situações o número necessário de vezes de forma a produzir alguma inferência estatística que permita validar uma das hipóteses formuladas.

No campo do controlo estatístico (SPC), pretende-se ainda salientar que, tal como foi dito, apesar de se ter a sensação de que o tema não foi muito desenvolvido, foram desenvolvidas ferramentas que poderão proporcionar a obtenção de bons resultados nesta área em trabalhos futuros tanto na vertente da integração EPC/SPC como na vertente do controlo estatístico multivariado com dados correlacionados. Na primeira vertente defende-se a ideia de que a utilização de modelos fiáveis em processos de simulação pode ser uma ferramenta de grande potencialidade na determinação da estratégia de controlo estatístico a integrar e do tipo de cartas a utilizar. Na segunda vertente, como também já foi referido, quanto mais próximo os resíduos do modelo estiverem de um processo ruído branco, mais eficaz será o controlo estatístico aplicado a esses resíduos e consequentemente ao processo (Figura 6-1).

No campo da integração do controlo estatístico com a engenharia de controlo, fizeram-se algumas abordagens, umas com mais hipóteses de sucesso que outras. Do ponto de vista teórico, a estratégia mais audaz de integração do controlo estatístico com a engenharia de controlo, seria monitorizar os parâmetros do modelo do processo através de um sistema de cartas Cuscore, para detectar alterações do processo em relação ao modelo interno do controlador. Esta teoria está bem desenvolvida para os modelos ARMA univariados (Box & Luceño, 1997) embora não se detectassem situações de aplicação real, talvez devido à sua complexidade. Contudo, tentar extrapolar a ideia para sistemas multivariados deverá tornar-se ainda mais complexo devido à grande quantidade de parâmetros envolvidos.

Há no entanto uma abordagem que se pensou fazer, caso o processo de estudo fosse o modelo base (Figura 8-1). A ideia base passa pelo seguinte; quando se está a falar de controlo cujo período de amostragem é da ordem das horas, trata-se de um controlo indexado à supervisão de processos. Com este período de amostragem faz todo o sentido utilizar o controlo estatístico para monitorizar variáveis de entrada não manipuláveis no sentido de detectar desvios em relação aos níveis esperados no sentido de ajustar as variáveis manipuláveis de modo a compensar esses desvios. Suponha-se que se tem um processo como o representado pelo diagrama de blocos da Figura 6-3 com duas variáveis de entrada, uma que não é manipulada e se supõe com nível constante $\boldsymbol{x_i}$, e outra que serve para controlar a saída do processo $\boldsymbol{x_c}$. Faz todo o sentido monitorizar a variável $\boldsymbol{x_i}$, recorrendo ao controlo estatístico (uma carta CUSUM ou EWMA por exemplo) de forma a detectar alterações nessa variável e a ajustar a variável $\boldsymbol{x_c}$ no sentido de compensar essa alteração (Figura 9-1).

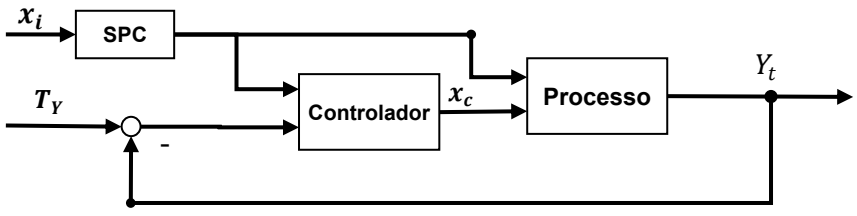


Figura 9-1 Exemplo de aplicação de carta de controlo à entrada do processo

Para explanar melhor este ponto de vista, de certa forma especulativo, suponha-se que a Figura 9-2 representa a função objectivo do controlador, que se pretende minimizar. Suponha-se que num determinado instante o nível da matéria-prima é de 0.2, e que a variável manipulável se encontra no nível 0.4, valor este determinado através de desenho de experiências, resposta em superfície ou outra metodologia. Se for detectada uma alteração de, por exemplo de 0.6 no nível da variável da característica da qualidade, faz todo o sentido alterar o nível da variável manipulada para zero no sentido de encontrar o ponto de funcionamento óptimo do processo, condicionado à variável de entrada x_i. Poderá sempre argumentar-se que essa característica poderia ser permanentemente medida, que seria o comportamento óptimo, mas nos processos industriais, e não só, existe sempre a componente rentabilidade associada, pelo que poderá ser mais económico ter um controlo por amostragem que uma medição continua. Por outro lado, por vezes as medições são bastantes dispendiosas, ou é mesmo necessário recorrer à destruição da amostra, pelo que também aí a integração do controlo estatístico com engenharia de controlo seria uma boa solução.

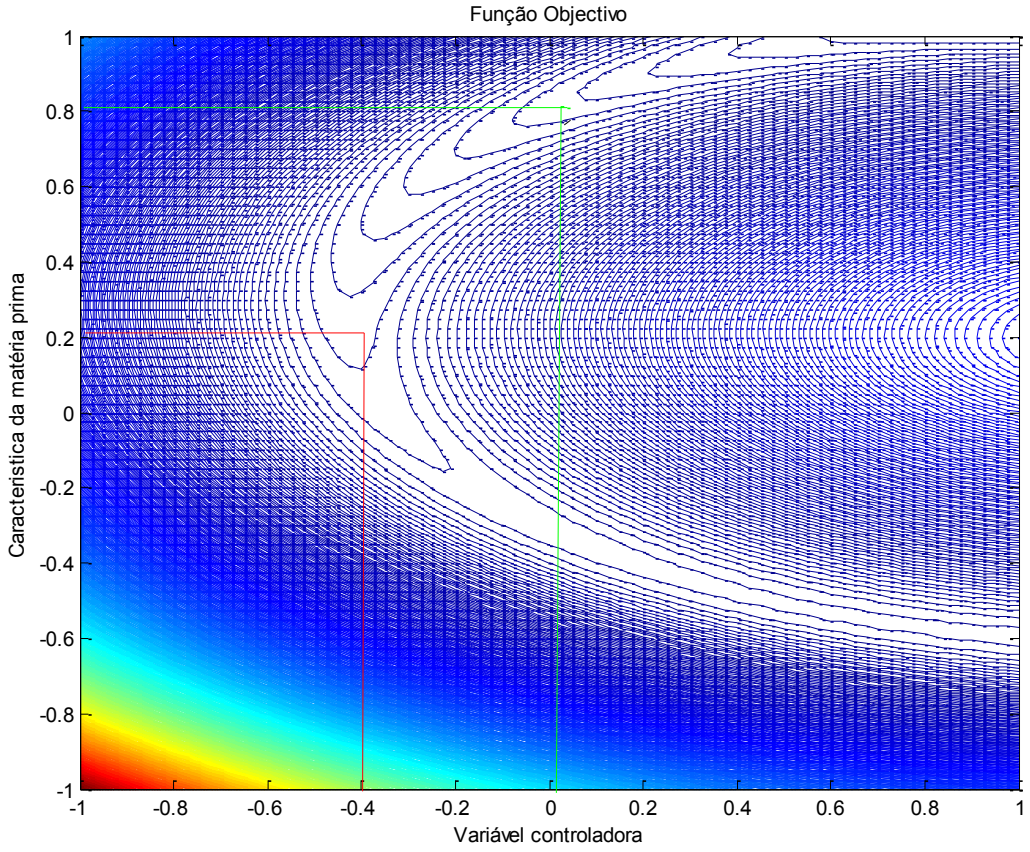


Figura 9-2 Exemplo de função objectivo de um sistema de controlo com duas variáveis de entrada

Quanto ao processo estudado, ficou a ideia geral de que as variáveis utilizadas como variáveis de controlo não eram as mais indicadas devido ao facto de também elas

apresentarem alguma dinâmica e de também elas dependerem de outras variáveis que não foram provavelmente aqui tidas em conta. Relembrando a representação em espaço de estados

$$\boldsymbol{x(k+1)} = \mathbf{A}\,\boldsymbol{x(k)} + \mathbf{B}\,\boldsymbol{u(k)}$$

$$\boldsymbol{y(k)} = \boldsymbol{Cx(k)} + \boldsymbol{D}\,\boldsymbol{u(k)}$$

em que $u(k)$ é o vector das variáveis de controlo (entrada), $x(k)$ é o vector das variáveis de estado e $y(k)$ é o vector das variáveis de saída, fica-se com a ideia de que faltam algumas variáveis pertencentes ao vector $u(k)$, e que algumas das variáveis utilizadas como pertencentes ao vector $u(k)$ deveriam pertencer efectivamente ao vector $x(k)$.

Ainda em relação ao modelo obtido, pode-se concluir que grande parte da variabilidade não explicada poderá dever-se à ausência no modelo desenvolvido de outras variáveis que não foram tidas em conta e com as quais a variável de saída estará fortemente correlacionada, como será o caso das variáveis referentes às características da matéria-prima. Há no entanto a salientar que, para os efeitos pretendidos, os dados (reais) disponibilizados serviram de forma satisfatória os objectivos.

Na conclusão sumária pode-se argumentar que, nos quatro grandes blocos que compõem este tema, muita coisa ficou por abordar, mas o trabalho mais urgente a fazer nesta área, é validar o que está feito através duma implementação prática em ambiente industrial.

A integração de lógica fuzzy nos sistemas de controlo e redes neuronais nos processo de identificação e/ou controlo são duas áreas que devem merecer alguma atenção em trabalhos futuros.

Referencias

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control* , AC-19, 716-723.

Alwan, L. C. (1988). Time-series modeling for statistical process control. *Journal of Business and Economic Statistics* , 6, 87-95.

Atienza, O. O. (1998). A SPC procedure for detecting level shifts of autocorrelated processes. *Journal of Quality Technology* , 30, 340-351.

Bartlett, M. S. (1946). On the theoretical specification and sampling properties of the autocorrelated time-series. *Journal of the Royal Statistics Society* , Ser. B, 8, 27-41.

Botto, M. A. (2007). Controlo de Sistemas. *Suplemento da Sebenta de Controlo de Sistemas* . IST - Lisboa.

Botto, M. A. (Setembro de 2007). Controlo Óptimo. *Sebenta de Controlo Óptimo* . IST - Lisboa.

Botto, M. A. (2007). Suplemento de Sebenta de Controlo de Sistemas. IST - Lisboa.

Box, G. E. (1992). Statistical process monitoring and feedback adjustment-a discussion. *Technometrics* , 34, 251-267.

Box, G. E., & Draper, N. R. (2007). *Response Surfaces, Mixtures, and Ridge Analyses.* JOHN WILEY & SONS, INC.

Box, G. E., & Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control.* Oakland, Calif.: Holden-Day.

Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2008). *Time Series Analysis: Forcasting snd Control, 4th ed.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Box, G., & Luceño, A. (1997). *Statistical Control By Monitoring and Feedback Adjustment.* JOHN WILEY & SONS, INC.

Capilla, C., Ferrer, A., Romero, R., & Hualda, A. (1999). Integration of statistical and engineering process control in a continuous polymerization process. *Technometrics* , 41, 14-28.

Chang, Z., Hao, D., & Baras, J. S. (2000). *http://hdl.handle.net/1903/6136* . Obtido em 10 de Janeiro de 2009, de Institute for Systems Research Technical Reports: http://hdl.handle.net/1903/6136

Chen, A. a. (2002). Design and performance evaluation analysis of the exponentially weighted moving average mean estimates for process subject to random step changes. *Technometrics* , 44, 379-389.

Chen, L., & Goldfarb, D. (August de 2006). Interior-point l2-penalty methods for nonlinearprogramming with global convergence properties. *MAthetmatical Programming* , pp. Vol. 108, 1-36.

del Castillo, E. (1996). A multivariate self-tuning controller for run-to-run process control under dhift and trend sisturbances. *IIE TRANSACTIONS* , vol. 28, no 12, 1011-1021.

Del Castillo, E. (1998). An adaptative run-to-run optimizing controller for linear and nonlinear semicndutor processes. *IEEE TRANSACTIONS ON SEMICONDUTOR MANUFACTORING* , pp. 285-295.

Del Castillo, E. (2002). *Statistical Process Adjustment for Quality Control.* New York: John Wiley & Sons, Inc.

Del Castillo, E., & Yeh, J.-Y. (1998). An adaptive run-to-run optimizing controller for linear and nonlinear semiconductor processes. *IEEE Transactions on Semiconductor Manufacturing* , Vol. 11, No. 2,pp. 285-295.

Grubb, F. E. (1954/1983). An optimal procedure for setting machines or adgusting processes. *Journal of Quality Technology* , 1983, 15(4): 186-189.

Hannan, E. J., & Rissanen, J. (1982). Recursive estimation of mixed autoregressive moving average order. *Biometrika* , 69, 81-94. Correction, 70, 303, 1983.

Isermann, R. (1989). *Digital Control Systems; Volume I; Fundamentals, Deterministic Control.* Springer-Verlag.

Jen, C. H., Jiang, B. C., & Fan, S.-K. S. (2004). General run-to-run (R2R) control framwork using self-tunning control for multiple-input multi-output (MIMO) process. *International Journal of Production Research* , 4249-4270.

Krishna B. Misra. (2008). *Handbook of Performability Engineering .* Springer.

Lowry, C. A., Woodall, W. H., Champ, C. W., & Rigdon, S. E. (1992). A Multivariate Exponentialy Weighted Moving Average Control Chart. *Technometrics* , Vol. 34, no 1, 46-53.

Luceño, A., Gonzalez, F. J., & Puig-Pey, J. (1996). Computing optimal adjustment schemes for the general tool wear problem. *Journal of Statistical Computation and Simulation* , 54, 87-113.

Lütkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis.* Berlin Heidelberg: Springer.

MacGregor, J. F. (1990). Discussion of `EWMA control schemes : properties and enhancement' by Lucas and Saccucci . *Technometrics* , 32, 23-26.

MacGregor, J. F. (1987). Interface between process control and on-line statistical process control. *American Institute of Chemical Engineers, Cast Newsletter* , 9-19.

Maciejowski, J. M. (2002). *Predictive Control With Constraints.* Essex, England: Pearson Education Limited.

Matos, A. S. (2006). Engenharia de Controlo do Processo e Controlo Estatístico da Qualidade: Metodologia de Integração Aplicada na Indústria da Pasta de Papel. *Tese de Doutoramento* . FCT/UNL – DEMI, Monte de Caparica.

Montgomery, D. C. (1994). Integrating statistical process control and engineering process control. *Journal of Quality Technology* , 26, 79-87.

Montgomery, D. C. (2001). *Introduction to Satistical Quality Control.* John Wiley & Sons, Inc.

Moyne, J., del Castillo, E., & Hurwitz, A. M. (2001). *Run-to-Run Control in Semicondutor Nanufacturing.* CRC press.

Muth, J. F. (1960). Optimal Properties of Expontially Weighted Forecasts of Time Series With Permanent and Transitory Components . *Journal of the American Statistical Association* , 55, 299-306.

Ogata, K. (1970). *Engenharia de Controle Moderno.* Brasil: Prentice-Hall.

Ogata, K. (1997). *Modern Control Engineering.* PRENTICE HALL.

Pan, R., & Del Castillo, H. (2003). E. Scheduling methods for statistical setup asjustment problems. *Intenational Journal of Production Research* , (41): 1467-1481.

Reinsel, G. C. (1997). *Elements of Multivariate Time Series Analysis.* New York: Springer-Verlag.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* , 6, 461-464.

Shu, L., Apley, D. W., & Tsung, F. (2002). Autocorrelated Process Monitoring Using Triggered Cuscore Charts. *Quality AND Reliablity Engeneering Interantional* , pp. 18, 411–421.

Sulo, P., & Vandevan, M. (1999). Optimal adjustment strategies for a process with run to run variation and 0-1 quality loss. *IIE Trans.* , 31:1135-1145.

Taguchi, G. (1987). *Systems of Experimental Design, Vol. 2.* Unipub, Kraus International Publications, White Plains, NY.

Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). Numerical Recipes: The Art of Scientific Computing. Cambrige, England.

Vander Weil, S. (1996). Monitoring processes that wander using integrated moving average models. *Technometrics* , 38, 139-151.

Vander Weil, S. T. (1992). Algorithmic statistical process control: concepts and an application. *Technometrics* , 34, 286-297.

Vanli, O. A., & Del Castillo, E. (NOVEMBER de 2007). Closed-Loop System Identification for Small Samples With Constraints. *TECHNOMETRICS* , VOL. 49, NO. 4, 382-394.

Wardell, D. G. (1994). Run-length distributions of special-cause control charts for correlated processes. *Technometrics* , 36, 3-17.

ANEXOS